Corpus Manual
German native speaker to non-native speaker long-term instant messaging dialogues
deL1L2IM
Release March 2015

Sviatlana Höhn

March 11, 2015

1 About this document

This user guide describes the data and the annotation for the collection of long-term instant messaging dialogues between native speakers of German and advanced learners of German as a foreign language.

Please sent all your comments, questions and suggestions concerning the dataset, the annotation and this manual to the author of this document *hoehn.sv @ gmail.com*.

2 Data

The data set has been collected in May-August 2012. This data set is part of a PhD project and is a side product of the project. The main focus of the PhD work was a study of main features of long-term interaction between language learners and language experts in order to obtain data-driven computational models for a conversational agent that helps to practice conversation in a foreign language.

2.1 Data Collection Setup

The participants of the data collection experiment were 9 advanced learners of German as a foreign language and 4 German native speakers. All of the learners were Russian native speakers and students of German at a Belorussian university. Native speakers are friends and colleagues of the organiser.

Each native speaker had 2 or 3 conversational partners. The organiser assigned the participants in pairs according to the time slots that the participants specified (when they have time to chat). Only the first appointment was arranged by the organiser. The only instruction was "just chat". The participants were expected to have a free conversation and to talk about whatever they want. The goal to produce 8 dialogues in total and to interact 4 weeks long 2 times per week was communicated to the participants.

The participants interacted using Google Talk infrastructure. A forwarding chatbot hosted on Google App Engine was used to collect the data on-line. Participants did not see each other directly, they sent the messages to the bot and the bot instantly forwarded the messages to the partner. All the chat logs were available immediately. The participants new that their talks were recorded. The participants agreed to publish the produced chat data prior to starting the interaction.

2.2 Data Selection

The participants produced 73 dialogues in total (8 dialogues by each of 6 pairs, 9 dialogues by each of 2 other pairs, and 7 dialogues by the 9th pair). Besides that the participants sometimes missed each other oder

forgot appointments. In these cases, the participants sent each other notifications ore excuses, but in some cases several days were between turns. The decision was made to include only full dialogue sessions in the final data set where participants met and talked. Each dialogue is between 20 and 45 minutes duration. The total size of the final dataset is 4548 messages that correspond to 236.302 text symbols. The message length ranges from 1 to 774 text symbols and of an average length of 58,5 symbols over all pairs.Table 1 summarises the statistics.

Metrics	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6	Pair 7	Pair 8	Pair 9	Total
MaxTL	335	405	774	313	414	277	637	460	232	
MinTL	1	1	2	1	2	2	2	2	1	
AvTL	62,98	72,40	105,13	38,13	86,80	38,99	42,23	48,85	31,38	58,54
# Turns	365	410	346	650	218	421	730	694	714	4548
# Symbols	22989	29683	36374	24784	18923	16413	30825	33903	22408	236302

Table 1: Corpus statistics: MaxTL - maximum turn length, MinTL - minimum turn length, AvTL - average turn length followed by the total number of turns and total number of symbols for each pair and in total for all pairs in the most right column.

2.3 Data Privacy

The data set was anonymised by replacing all the user names by placeholders. All information that could disclose the identity of the participants (like links to Facebook profiles or email addresses) were replaced by placeholders. The information about the language proficiency of the participant is encoded in the placeholder: NATIVE + Nr. specifies a native speaker, LEARNER + Nr. specifies a learner. LN and FN specify the placeholders for last name and first name of the participant.

Example: NATIVE02_FN is a placeholder for the first name of the native speaker number 01.

3 Annotation

It was not the goal of the work to create a comprehensive annotation schema for chat data. However, particular phenomena have been annotated in the dataset for the needs of the PhD project. Only annotation of those phenomena is contained in the present version of the corpus and in the TEI-P5 customisation.

The annotation was performed by two independent annotators in August-September 2013. The first annotator was German non-native speaker with a near-native fluency in German and Russian as native language with strong professional background in linguistics, natural language processing and language teaching. The second annotator was German native speaker with no professional background in linguistics but with strong knowledge of the German language. In August 2014 and in February 2015, the annotation was rechecked by the first annotator, several annotation errors and ambiguous cases were corrected.

3.1 TEI-P5 Modules and Customisation

TEI-P5 standard already contains several customisations for spoken interaction data, poetry, linguistic corpora and drama. Chat or instant messaging data have similar features with some of them, but none of the existing customisations could be used without modification and with the appropriate semantics of the tags. Therefore, a decision was taken to create a new customised annotation scheme using the existing TEI customisation for linguistic corpora as a basis. This allows for continuous extension of the annotation of linguistic phenomena in the dataset.

For the purpose of the related PhD project only annotation of repair sequences in chat has been performed, the TEI schema was customised according to the requirements annotation (see Section 3.4 below).

The corpus is provided as a set of 10 files: one root file containing the description of the corpus and information about participants (TEI header), and 9 files with chat logs produced by 9 pairs of participants, one file per pair. Each file containing chat logs includes all dialogues produced by one pair of participants. The root file contains links to each of those files.

3.2 Text Replacements

In general, the original spelling and textual symbols used are kept as produced by the used. However, there are a few exceptions made for the purpose of storing the data in XML format and data analysis. All the replacements are summarised in Table 2.

Original	Replaced by
&	&amp;
All posted hyperlinks	HYPERLINK
Facebook ID	FACEBOOK_ID_{LXX,NXX}
Email address	EMAIL

Table 2: Text replacements

3.3 Chat Structure in TEI-P5 XML

The goal of this annotation was to provide a TEI-P5 conform encoding for chat data. Two new tags were introduced:

- `<message>` contains the text of one instant message produced by a chat participant OR more than one non-empty message line (`<ml>` tag). Message lines were introduces for the cases where chat participants insert line breaks in their messages. Different lines may relate to different previous messages of the partners, and need to be linked separately. Important attributes of a message are sender, timestamp and id. The sender is specified by the standard TEI attribute `who` and is linked to a chat participant listed in the root file. The `timestamp` attribute specifies the server time when the message arrived at the server (time zone GMT+0, needs to be recalculated to determine factual time in the time zones of the chat participants - Germany and Belarus).
- `<ml>` contains a message line if and only if the sender of the message inserted breaklines in the message.

The corresponding schema is contained in the file `tei_corpus_chat.rng` which is provided with the corpus.

3.4 Chat Log Annotation - Repair

For the purpose of the PhD project, the data have been annotated according to the two classifications:

1. Corrective Feedback as explained by Lyster, Sato Saito (2013). The sequences containing CF usually consist of three types of interactional moves described in classroom research literature:error, correction, uptake.
2. Meaning Negotiation as introduced by Gass and Varonis (1985). According to this model, a MN sequence is composed of 4 moves: trouble source, indicator, response, reaction to response. The messages were labelled according to these types of moves.

All moves in these sequences may consist of several turns (messages). The TEI `note` tag us used for explanations of some complicated cases of annotation (ambiguities, complex sequences).

All of these sequences are repair sequences from the point of view of Conversation Analysis. However, in the beginning of the related PhD project the data analysis was influenced by the language classroom research and Second Language Acquisition theory.

3.4.1 Corrective Feedback

Types of Corrective Feedback (CF) adopted for the PhD work are:

- conversational recast,
- repetition,
- clarification request,
- explicit correction,
- explicit with metalinguistic explanation (MLE),
- didactic recast,
- metalinguistic clue,
- elicitation,
- paralinguistic signal.

The types of the corrections are explained in the article by Lyster, Sato, Saito (2013). Not all of them occur in the dataset because of the text based nature of the chat and due to an informal conversation contrasting to a teacher-fronted classroom where the original classification was obtained from. Metalinguistic clues, elicitations, repetitions and paralinguistic signals were not found in CF-sequences in this dataset.

An example of an annotated CF-sequence is provided below:

```
<im:message xml:id="L06N0320120710-281" who="deL1L2IM-root.xml#L06"
timestamp="2012-07-10T08:44:09"> man kann versuchen </im:message>
<im:message xml:id="L06N0320120710-282" who="deL1L2IM-root.xml#N03"
timestamp="2012-07-10T08:44:29"> [[wir koennen es versuchen]] </im:message>
<im:message xml:id="L06N0320120710-283" who="deL1L2IM-root.xml#N03"
timestamp="2012-07-10T08:44:32"> :-) </im:message>
<im:cfseq>
<im:cfturn turntype="ts" corresp="L06N0320120710-281"/>
<im:cfturn turntype="cf" corresp="L06N0320120710-282"
cftype="explicit_correction"/>
</im:cfseq>
```

No uptake was produced in this sequence.

3.4.2 Incomplete Error Annotation

The error annotation has been performed in place (the error is tagged where it occurs). If an item is missing (missing main verb or missing prefix), the content of the tag is empty. The following error types have been annotated in the corpus (all occurrences of them, not only the corrected ones).

1. Morpho-syntactic errors:
   a) Missing main verb in Futurum 1: missing_main_verb_futur1
   b) Wrong word order in a sentence. Only the following types have been annotated:

- Wrong position of the main verb in the main or subordinate close, or missing main verb.
  Possible types:
  missing_finiite_verb_main_clause
  missing_finiite_verb_subordinate_clause
  position_finiite_verb_main_clause
  position_finiite_verb_subordinate_clause
- Wrong posistion of or missing separable prefix in verbs.
  Possible types:
  missing_verb_prefix_main_clause
  missing_verb_prefix_subordinate_clause
  position_verb_prefix_main_clause
  position_verb_prefix_subordinate_clause.

2. Lexical erros: only lexical errors in collocations. Definition, classification and types for locations: s. Paper "Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora". Examples of collocations: frei haben, leicht fallen, Fliege machen.
   Possible types:
   substitution | creation | synthesis | analysis | different_sense
   Possible locations:
   base | collocate | collocation

```
<im:error errtype="position\_finite\_verb\_subordinate\_clause"
target="man kann darüber recherchieren" corrected="NO">
man darüber rescherschieren kann</im:error>.
```

NOTE: Wie oft, der oben markierte Fehler (man darüber rescherschieren kann) kann auf unterschiedliche Weise korrigiert werden (mehrere korrekte Target-Formen). Hier habe ich mich meistens für eine der möglichen korrekten Formen entschieden. Eine vollständige Fehler- und Korrekturannotation war nicht das Ziel dieser Arbeit. Es gibt aber eine Möglichkeit, mehrere Varianten zu annotieren. Dafür kann das Alternative-Tag verwendet werden. Ein Beispiel für eine solche Annotation ist hier angegeben:

```
<im:message xml:id="L07N0320120716-602" who="deL1L2IM-root.xml#L07"
 timestamp="2012-07-16T19:29:15"> [content hidden for the example]
 und das wichtigste ich müsse so spät wie möglich heiraten,
<altGrp>
 <alt>
  <im:error errtype="substitution" location="collocate"
   target="sonst geht alles kaputt" corrected="NO"/>
  </alt>
 <alt>
  <im:error errtype="missing_main_verb_futur1"
   target="sonst wird alles kaputt gehen" corrected="NO"/>
 </alt>
sonst wird alles kaputt
</altGrp>
 ...)))))) xD </im:message>
```

3.4.3 Meaning Negotiation

According to the model of a Meaning Negotiation (MN) sequence suggested by Varonis & Gass (1985), the classification of the dialogue moves includes the following classes:

- Trouble-source: the problematic item that is not clear and needs a clarification;
- An indicator that something previously said is not clear;
- A response to the indicator (normally a clarification or an explanation);
- A reaction to response (for example an acceptance or a rejection of a term).

This model is very simple, MN sequences can be very complex, an indicator for example can be also a trouble source and trigger a nested MN sequence or a cascade of MN sequences. To tag such sequences I still used this basic model and handled each nested or cascading sequence as part of a large sequence.

- `mnseq`: Contains a Meaning Negotiation sequence.
- `mnturn`: Contains a reference to a turn that is part of a specific `mnseq`. Turn types must be specified in the attribute `turntype` wich contains the type of the turn that it part of this Meaning Negotiation sequence. Typical turn types are trouble-source (ts), indicator (ind), response (resp), reaction to response (rr). Types are not predefined in the schema because different types not fitting in the basic classification are possible. Cascading and nested MN sequences can be also part of `mnseq` (see turns from L08N0420120531-226 to L08N0420120531-237 of the corpus and their annotation).

Example of a Meaning Negotiation sequence:

```
<im:message xml:id="L01N0120120618-268" who="deL1L2IM-root.xml#N01"
   timestamp="2012-06-18T21:02:44"> Das schaffst du mit links :-)
</im:message>
<im:message xml:id="L01N0120120618-269" who="deL1L2IM-root.xml#N01"
   timestamp="2012-06-18T21:02:55"> (Verstehst du die Bedeutung?)
</im:message>
<im:message xml:id="L01N0120120618-270" who="deL1L2IM-root.xml#L01"
   timestamp="2012-06-18T21:03:42"> Und was bedeutet das?
</im:message>
<im:message xml:id="L01N0120120618-271" who="deL1L2IM-root.xml#N01"
   timestamp="2012-06-18T21:04:37"> mit links schaffen/erledigen/
   erreichen = ohne Probleme schaffen/erledigen/erreichen
</im:message>
<im:message xml:id="L01N0120120618-272" who="deL1L2IM-root.xml#L01"
   timestamp="2012-06-18T21:05:46"> Vielen Dank für Erklärung :)
</im:message>
<im:message xml:id="L01N0120120618-273" who="deL1L2IM-root.xml#N01"
   timestamp="2012-06-18T21:06:36"> Kein Problem :-)
</im:message>

<im:mnseq>
  <im:mnturn turntype="ts" corresp="L01N0120120618-268"/>
  <im:mnturn turntype="ind" corresp="L01N0120120618-269"/>
  <im:mnturn turntype="ind" corresp="L01N0120120618-270"/>
  <im:mnturn turntype="resp" corresp="L01N0120120618-271"/>
  <im:mnturn turntype="rr" corresp="L01N0120120618-272"/>
</im:mnseq>
```

References

Roy Lyster, Kazuya Saito and Masatoshi Sato (2013). Oral corrective feedback in second language classrooms. Language Teaching, 46, pp 140, doi:10.1017/S0261444812000365

Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, and Sabela Prieto González. (2010) Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In: LREC-2010.
Evangeline Marlos Varonis and Susan Gass (1985). Non-native/non-native conversations: A model for negotiation of meaning. Applied Linguistics, 6(1):71–90.