Joshgun Sirajzade

Korpusbasierte Untersuchung der Wortbildungsaffixe im Luxemburgischen. Technische Herausforderungen und linguistische Analyse am Beispiel der Produktivität

Abstract: The project WBLUX (Wortbildung des moselfränkisch-luxemburgischen Raumes) at the University of Luxembourg aims at the investigation of Luxembourgish word formation through different text sorts and genres. In order to achieve this goal the compilation of an annotated corpus is needed. This article gives an example for benefits of using a corpus with annotations like parts of speech, lemmata and word formation affixes in the analysis of productivity of some selected word formation affixes of Luxembourgish. Then it describes how one can achieve such a corpus from a technical point of view. This includes the choice of corpus format, of a database platform and the designing of programs needed for the annotation process of word formation itself. This article also suggests new corpus linguistic approaches for research of word formation like analyzing the usage of word formation bases in the entire corpus or performing context analysis in order to determine semantical functions of each suffix.

1. Einleitung

Korpuslinguistische Methoden sind nicht nur ein fester Bestandteil der heutigen linguistischen Forschung, sondern auch eine notwendige Voraussetzung für die zuverlässige, empirische und nachhaltige Untersuchung einer Sprache. Somit sind sie auch nicht aus der Wortbildungsforschung des Luxemburgischen wegzudenken. Sie erlauben sowohl eine strukturierte Organisation der Daten dank Technologien wie Datenbanken, die zum Auffinden und Speichern von gesuchten Sprachphänomenen sehr hilfreich sein können, als auch die schnellere Durchführung von formalen und stochastischen Analysen. Dabei kann man außer der korpuslinguistischen Theorie selbst auch auf die Erkenntnisse der Informatik sowie Linguistik und insbesondere der einzelnen benachbarten interdisziplinären Fachrichtungen wie Computerlinguistik, Quantitative Linguistik sowie Digital Humanities zurückgreifen. Die Erforschung der luxemburgischen Sprache, die oft als jüngste germanische Sprache bezeichnet wird, hat inzwischen eine Tradition. Eine Übersicht findet sich bei Moulin (2004). Dennoch sind grammatische Beschreibungen und Analysen des Luxemburgischen im Vergleich zu Untersuchungen der benachbarten Sprachen immer noch rar (Gilles 2006). Besonders fällt dies im Bereich der korpuslinguistischen Analysen auf (Sirajzade 2013: 5). Obwohl

bereits elektronische Ressourcen und Datenbanken für die Untersuchung des Luxemburgischen existieren (Sirajzade 2013: 5), gibt es hier einen Bedarf nach einem nationalen Sprachkorpus, das etwa vergleichbar wäre mit dem British National Korpus[1] oder dem Deutsche Referenzkorpus[2], um einige von vielen Beispielen weltweit zu nennen. Solch ein Korpus für das Luxemburgische würde eine repräsentative Untersuchung ermöglichen und die Ergebnisse der linguistischen Analysen vergleichbar machen. Das Projekt „Die Wortbildung des moselfränkisch-luxemburgischen Raumes (WBLUX)" am Institut für luxemburgische Sprache und Literatur der Universität Luxemburg hat sich zum Ziel gesetzt, die luxemburgische Wortbildung flächendeckend über verschiedene Stile und Textsorten zu untersuchen. Zu diesem Zweck mussten technische Hürden überwunden werden, wie die Aufbereitung des Korpus aus bestehenden elektronischen Texten sowie die Suche und Annotation der Wortbildungsaffixe. Dieser Umstand eröffnete zur gleichen Zeit eine Möglichkeit, die aktuellsten Standards und Technologien in der Korpuslinguistik zu untersuchen und einzusetzen. Anschließend mussten neuste Erkenntnisse und Methoden aus der Wortbildungsforschung elaboriert und angewandt werden. Dieser Aufsatz fungiert als ein Bericht, wie auf der einen Seite bereits existierende elektronische Texte des Luxemburgischen zu einem annotierten Korpus zur Untersuchung der Wortbildung aufbereitet werden können. Auf der anderen Seite werden statistische Möglichkeiten vorgestellt und diskutiert, mit deren Hilfe die Wortbildung in einem Korpus untersucht werden kann – konkret eine formale Analyse der Produktivität der ausgewählten Wortbildungssuffixe des Adjektivs und Substantivs im Luxemburgischen. Es soll gezeigt werden, wie sich die Entwicklungstendenzen der Wortbildungssuffixe im Luxemburgischen darstellen. Dabei werden Kategorien wie die realisierte und expandierende Produktivität diskutiert und konkrete Beispiele für die Konkurrenz der einzelnen analysierten Suffixe beschrieben. Diese Untersuchungen fanden im Rahmen des WBLUX Projekts an der Universität Luxemburg statt.

1.1. Forschungsstand

Die bisherige Forschungsarbeit, die für die Anwendung von korpuslinguistischen Methoden in der Wortbildungsforschung relevant sein kann, lässt sich in zwei Gruppen unterteilen: Die Erkenntnisse hinsichtlich der allgemeinen Algorithmen und Werkzeuge in der Linguistik einerseits, und andererseits die Erkenntnisse der Wortbildungsforschung selbst. Eine gute Übersicht darüber findet sich im HSK-Band ‚Word-formation' (Müller 2015: 2333–2372), in dem die Werkzeuge und Algorithmen aufgrund des zugrunde liegenden

1 http://www.natcorp.ox.ac.uk/, zuletzt geprüft am 29.11.2017.
2 http://www1.ids-mannheim.de/kl/projekte/korpora/, zuletzt geprüft am 29.11.2017.

empirischen Materials in drei einzelnen Aufsätzen behandelt werden: 1. Dictionaries (Belentschikow 2015); 2. Corpora (Heid 2015) und 3. Internet (Dal und Namer 2015). Den Nutzen der Wörterbücher sehen die Autoren in den Informationen zur Wortbildung, die diese zur Verfügung stellen, zum einen in den einzelnen Wortartikeln der einsprachigen Wörterbücher. Zum anderen existieren spezialisierte Wörterbücher, insbesondere die sogenannten „morpheme and word-formation dictionaries", die bereits eine Liste von Affixen mit ausführlichen Informationen zu Wortbildung enthalten (Belentschikow 2015: 2345). Hinsichtlich der Wortbildung des Luxemburgischen findet sich leider kein solches Wörterbuch. Eine besondere Rolle spielt im HSK-Band der Aufsatz ‚Corpora' (Heid 2015). Hier werden nicht nur die wichtigsten Eigenschaften eines Korpus wie Repräsentativität, Zusammenstellung und Größe diskutiert, sondern es werden auch wichtige Strategien in der Wortbildungsforschung wie die Verwendung der vorhandenen Annotationen (Part-of-speech (POS), Lemmata etc.) bei der Suche nach Wortbildungselementen, deren automatisierte Zerlegung, die Festlegung der Produktivität durch Frequenzanalysen und die Untersuchung der Funktionsklassen durch Kontexte aufgezeigt (Heid 2015: 2354–2371). Einiges an Arbeit wurde auch bereits geleistet, um Werkzeuge zu entwickeln, die eine automatisierte morphologische Zerlegung und damit auch die Zerlegung der Wortbildungselemente gewährleisten können. Was die in der Korpuslinguistik etablierten Werkzeuge angeht, existiert bereits ein POS-Tagger für das Luxemburgische (Sirajzade 2012: 264–280). Darüber hinaus existieren andere korpuslinguistische Werkzeuge, die zwar nicht explizit für das Luxemburgische implementiert sind, jedoch stochastische Mittel verwenden und im Prinzip sprachunabhängig sind. Eines davon ist Morfessor[3], hauptsächlich entwickelt für das Finnische, das eine relativ komplexe morphologische Struktur aufweist (Creutz und Lagus 2002; Kohonen et al. 2010). Eine andere Herangehensweise ist die semiautomatische morphologische Zerlegung, die meist auf einem Sprachmodell basiert. Getestet am Material des Arabischen ergab diese Herangehensweise einen Korrektheitsgrad von 97% (Lee et al. 2003: 399–406). Einen deutlich klassischeren Weg geht das Projekt MorphoDiTa[4], das hauptsächlich aus einem POS-Tagger und einem morphologischen Wörterbuch besteht. Die beiden Teile des Programms können trainiert werden (Straková et al. 2014). Es gibt bereits für das Tschechische und Englische trainierte Modelle, die man verwenden kann. Jedoch betreffen alle diese Arbeiten die gesamte Morphologie und beschäftigen sich nicht zielgerichtet mit der Wortbildung. Bisher wurde viel Forschung hinsichtlich der morphologischen Produktivität und ihrer quantitativen Modellierung betrieben (Baayen 2009; Pustylnikov und

3 mittlerweile in der Version 2.0. http://morpho.aalto.fi/projects/morpho/, zuletzt geprüft am 29.11.2017

4 http://ufal.mff.cuni.cz/morphodita, zuletzt geprüft am 29.11.2017

Schneider-Wiejowski 2010; Schneider-Wiejowski 2011). Eine Beschreibung der Untersuchungsgeschichte der luxemburgischen Wortbildung findet sich bei Sirajzade (2013: 57–60). Außer den hier ausgeführten klassischen Grammatiken wäre die Arbeit von Lulling (2002) zu nennen, der die Kreativität der luxemburgischen Wortbildung anhand eines selbst kompilierten Korpus untersucht.

1.2. Das Korpus

Die Korpusgrundlage für die Untersuchung der luxemburgischen Wortbildung besteht aus zwei Teilkorpora: 1. Ein ausgewogenes, repräsentatives Kernkorpus, bestehend aus ca. einer Mio. Token. 2. Ein Erweiterungskorpus für weitere, meist stochastische Analysen, bestehend aus ca. 19 Mio. Token. Insgesamt bestehen beide Teilkorpora aus knapp unter 20 Mio. Token. Hierzu gehören Parlamentsreden (Chambre des Députés) mit ca. 10 Mio. Token in 293 Dokumenten, RTL News (Radio Télé Luxembourg) Nachrichten mit ca. 5 Mio. Token in 109 Dokumenten, luxemburgische Literatur vertreten mit ca. 2 Mio. Token in 54 Dokumenten sowie Interviews der Uni Luxemburg mit ca. 0,5 Mio. Token in 49 Dokumenten.

2. Statistische Analyse der Produktivität

2.1. Festlegung der Produktivität

Die Produktivität der Wortbildungselemente ist eine der wichtigsten Eigenschaften und zeichnet den Charakter und die Dynamik einer Sprache aus. So thematisiert Heid (2015: 2361) die Produktivität im HSK-Band „Word-formation“ ausführlich und beruft sich bei der mathematischen Beschreibung des Problems hauptsächlich auf die Arbeiten von Baayen. Baayen (2009) versucht, das Phänomen der Produktivität theoretisch und mathematisch zu formulieren, allerdings im Hinblick auf die gesamte Morphologie. In einfachster Form wird die Produktivität einer morphologischen Kategorie C geschätzt mit der Typezahl V(C, N), wobei N die Gesamtanzahl der Tokens in einem Korpus ist (Baayen 2009: 902). Baayen nennt dies die realisierte Produktivität. Durch die Miteinbeziehung der Types einer Kategorie spiegelt diese Maßgröße deren Möglichkeit zur lexikalischen Vielfalt wieder (Sirajzade 2013: 50). Laut Baayen ist die realisierte Produktivität der aktuelle Zustand in einer Sprache, sie repräsentiert jedoch nicht die Tendenzen einer Sprache. Dafür ist die expandierende Produktivität (‚expanding productivity‘) verantwortlich. Sie bezieht Hapaxlegomena einer Kategorie in die Berechnung ein, also die Anzahl der Wörter V(1, C, N) in einer Kategorie C, die in einem Korpus von der Größe N nur einmal auftauchen. Wenn V(1, N) die totale Anzahl der Hapaxlegomena in einem Korpus darstellt, dann schätzt das Verhältnis P* = V(1, C, N) / V(1, N) den Beitrag einer Kategorie zur Erweiterung des Lexikons einer Sprache. Eine

weitere Formel von Baayen, genannt die potenzielle Produktivität P* = V(1, C, N) / N(C), setzt die Hapaxlegomena ins Verhältnis zu der Gesamtanzahl der Tokens in einem Korpus. Die potenzielle Produktivität beschreibt das Wachstum des Lexikons einer grammatischen Kategorie (Baayen 2009: 902).

2.2. Produktivität bei der Adjektivbildung

Während *-esch* als adjektivbildendes Suffix 4 273 Mal in 737 Types im Kernkorpus (1 000 000 Tokens) vorkommt, lässt sich *-eg* als adjektivbildendes Suffix in 4 341 Tokens bestehend aus 489 Types finden. Im Verhältnis zur Gesamtanzahl der Tokens im Kernkorpus ergibt sich, dass die realisierte Produktivität von *-eg* 489/1 000 000 ≈ 0,000489 beträgt und bei *-esch* 737/1 000 000 ≈ 0,000737. Zum Suffix *-eg* gibt es 211 Hapaxlegomena im Kernkorpus, d.h. die expandierende Produktivität für *-eg* beträgt hier 211/16 372 ≈ 0,0129, während sie bei *-esch* mit 346 Hapaxlegomena 346/16 372 ≈ 0,0211 beträgt. Laut beiden Messungen ist *-esch* bei der Adjektivbildung im Luxemburgischen demnach produktiver als *-eg*, obwohl die Tokenanzahl bei *-eg* höher ist. Das wird offensichtlicher, je mehr Daten man in die Analyse miteinbezieht. Im Gesamtkorpus gibt es ca. 50 000 Tokens zu *-eg*, zu *-esch* lediglich ca. 45 000. Das bedeutet, dass die mit *-eg* gebildeten Adjektive in der sprachlichen Realisierung häufiger verwendet werden, aber *-esch* viel mehr lexikalische Vielfalt bei den Basen besitzt, mit denen es Verbindungen eingeht.

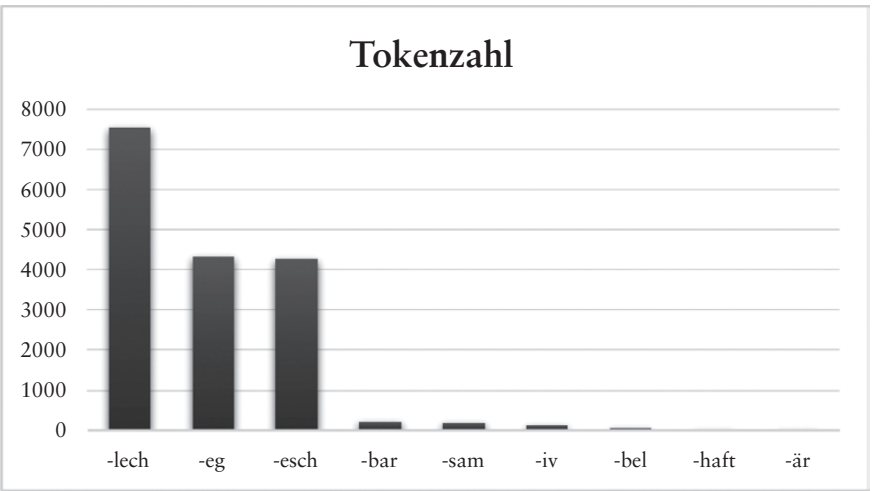
Betrachtet man die Zahl der Types bei dem Suffix *-lech,* das tokenmäßig häufigste adjektivbildende Suffix im Kernkorpus (7 555), so lässt sich feststellen, dass das Suffix *-lech* auch hinter *-esch* steht (556). Die Zahl der Hapaxlegomena ist mit 211 auch relativ niedrig. So kommt man auf eine realisierte Produktivität von 556/ 1 000 000 ≈ 0,000 556, wobei die expandierende Produktivität dabei 211/16 372 ≈ 0,0129 beträgt, ähnlich wie bei dem Suffix *-eg*. In der folgenden Tabelle finden sich die Token-, Type- und Hapaxlegomenon-Frequenzen der häufigsten adjektivbildenden Suffixe.

Tab. 1: Adjektivbildende Suffixe im Luxemburgischen.

Suffix	*POS*	*Tokenzahl*	*Typezahl*	*Hapaxlegomena*
-lech	ADJ	7555	556	211
-eg	ADJ	4341	489	211
-esch	ADJ	4273	737	346
-bar	ADJ	231	92	51
-sam	ADJ	183	18	10
-iv	ADJ	115	12	0
-bel	ADJ	59	28	5
-haft	ADJ	52	22	11
-är	ADJ	2	1	0

Außer diesen drei produktivsten adjektivbildenden Suffixen, kann man eine andere Gruppe mit geringerer Tokenanzahl unterscheiden. Zu dieser Gruppe gehören *-bar, -sam, -iv, -bel, -haft*. Zu *-bar* gibt es im Kernkorpus 232 Tokens und 94 Types, wovon 51 Hapaxlegomena sind. Bei dem Suffix *-sam*, welches im Kernkorpus 183 Mal in 18 Types vorkommt, gibt es nur 10 Hapaxlegomena. Das Suffix *-iv* verhält sich ähnlich; für 115 Tokens gibt es 12 Types, wobei es hier keine Hapaxlegomena gibt. Das Element *-bel* wie bei *honorabel* kommt als Token 59 Mal und als Type 28 Mal vor. Hapaxlegomena dazu gibt es nur in fünf Fällen. Zu *-haft* lassen sich 52 Tokens, 22 Types und 11 Hapaxlegomena finden, was zu folgenden Ergebnissen führt: Die realisierte Produktivität für das Suffix *-bar* beträgt 94/1 000 000 ≈ 0,000 094, für *-sam* 18/1 000 000 ≈ 0,000 018, für *-iv* 12/1 000 000 ≈ 0,000 012, für *-bel* 28/1 000 000 ≈ 0,000 028 und für *-haft* 22/1 000 000 ≈ 0,000 022. Also haben in dieser Gruppe die Suffixe *-bar* und *-bel* die höchsten Werte für die realisierte Produktivität. Die expandierende Produktivität für *-bar* ist mit 92/16 372 ≈ 0,0056 jedoch deutlich höher als bei *-bel* mit 5/16 372 ≈ 0,00031.

Abb. 1: Tokenzahl der adjektivbildenden Suffixe im Luxemburgischen.



Hapaxlegomena sind eine wichtige Eigenschaft eines Korpus, sie werden neben und im Gegensatz zu der Token-Type-Ratio, auch in der sogenannten Hapaxlegomena-Token-Ratio verwendet (Ali und Hussein 2014). Pustylnikov und Schneider-Wiejowski (2010: 13) schlagen in ihrer Untersuchung der Wortbildungssuffixe *-nis, -ung, -er* uns *-heit/-keit* im Deutschen vor, die komplette Distribution der Types eines Wortbildungselements zu berücksichtigen, allerdings erhalten sie bessere Ergebnisse mit dem Verfahren, das Hapaxlegomena berücksichtigt. Die Anzahl der Hapaxlegomena in einem

Korpus ist jedoch von dessen Größe abhängig. Im Gesamtkorpus (ca. 20 Mio. Tokens) ist zu *-esch* 1370 Hapaxlegomena zu finden, zu *-eg* nur 736, was den Unterschied in der Produktivität noch deutlicher macht, da die expandierende Produktivität von *-esch* 1 370/ 85 424 ≈ 0,016 deutlich höher ist als bei *-eg* 736/85 424 ≈ 0,0086.

Bei der Analyse der Hapaxlegomena im Falle von *-esch* und *-eg* lässt sich erkennen, dass *-esch* eher Bildungen mit Fremdwörtern romanisch oder gar lateinischen Ursprungs eingeht als *-eg*, z. B. *hygienesch* ‚hygienisch', *mikrobiologesch* ‚mikrobiologisch', *organisatoresch* ‚organisatorisch', *olympesch* ‚olympisch', *diabolesch* ‚diabolisch', *didaktesch* ‚didaktisch', *dynamesch* ‚dynamisch', *pädagogesch* ‚pädagogisch', *thematesch* ‚thematisch', *philosophesch* ‚philosophisch', *statistitesch* ‚statistisch', *ekonometresch* ‚ökonometrisch', *logistesch* ‚logistisch' etc. Es gibt aber auch ursprünglich luxemburgische Wörter, die mit *-esch* gebildet werden können, wie z. B. *ënnerierdesch* ‚unterirdisch', *stiermesch* ‚stürmisch', *auslännesch* ‚ausländisch'. Das Suffix kann außerdem nahezu mit allen Länder-, Orts- und Flurnamen kombiniert werden: *dänesch* ‚dänisch', *irlännesch* ‚irländisch', *australesch* ‚australisch', *éisträichesch* ‚österreichisch', *südeuropäesch* ‚südeuropäisch' etc. Das zeigt die Dynamik und das Potenzial des Suffixes. *-eg* hingegen bildet mehrheitlich Wörter von luxemburgischen bzw. germanischen Stämmen: *stéchhalteg* ‚stichhaltig', *réckfälleg* ‚rückfällig', *zweetrangeg* ‚zweitrangig', *gemengnëtzeg* ‚gemeinnützig', *zäitwëlleg* ‚zeitweilig', *eekeleg* ‚eklig', *flësseg* ‚flüssig', *gehéiereg* ‚gehörig'. Das Suffix *-eg* ist nicht offen für Wörter romanischen bzw. lateinischen Ursprungs, ist dennoch eines der produktivsten adjektivbildenden Suffixe im Luxemburgischen. Die Tatsache, dass man *-eg* im Korpus in Wörtern mit griechisch-lateinischen Ursprung findet, wie z. B. *drasteg*, welches im Korpus zwei Mal registriert wurde (RTL_News_National_06.2 und Parlament 2003–2004_09), ist auf einen Rechtschreibfehler zurückzuführen. Die richtige Schreibweise ist *drastesch* ‚drastisch'.

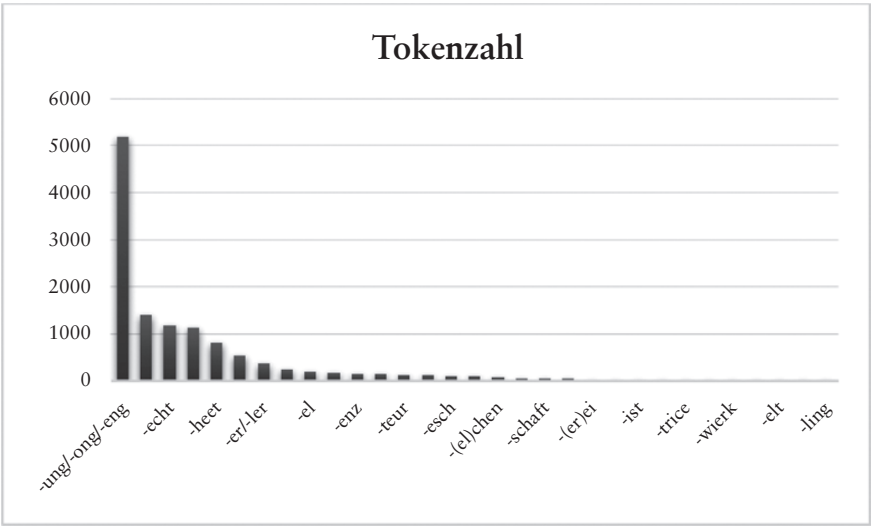
Was die Adjektive angeht, die mit *-bar* gebildet werden, ist deutlich der Einfluss des benachbarten Deutschen zu sehen, obwohl es aufgrund der Tatsache, dass beide Sprachen verwandt sind, schwer ist, diesen Einfluss ganz genau in Zahlen zu fassen. Es gibt im Kernkorpus Wörter wie *absehbar, wunderbar, furchtbar, fruchtbar* die deutlich deutschen Ursprungs sind, jedoch auch andere Wörter wie *bemierkbar* ‚bemerkbar', *ëmsetzbar* ‚umsetzbar', *virstellbar* ‚vorstellbar', *notzbar* ‚nutzbar', deren Ursprung strittig sein könnte (gebildet nach dem deutschen Muster) und Wörter, die sehr Luxemburgisch erscheinen wie *cumuléierbar* ‚kumulierbar', *uklobar* ‚anklagbar', *drobar* ‚tragbar'. Auf jeden Fall scheint das Suffix im Luxemburgischen in der zweiten Gruppe der Suffixe mit weniger Tokens am produktivsten zu sein, aufgrund der lexikalischen Vielfalt, die man auf den ersten Blick feststellen kann. Zahlen aus dem Korpus und die Berechnungen der realisierten und expandierenden Produktivität bestätigen

dies nochmals. *-bar* konkurriert jedoch mit dem Suffix *-bel*, das laut Produktivitätsmessungen zwar weniger produktiv ist als *-bar*, dennoch kommt es hinter *-bar* und ist produktiver als alle übrigen Suffixe in der zweiten Gruppe. *-bel* ist in Adjektiven romanischen Ursprungs zu finden, wie z. B. *applicabel* ‚anwendbar', *recevabel* ‚zulässig', *flexibel* ‚flexibel', *indispensabel* ‚indispensabel', *formidabel* ‚großartig' etc. Man könnte meinen, beide Suffixe haben ihre eigene Nische, führt man allerdings eine Basisanalyse durch, die als Verfahren später diskutiert wird, dann stellt man fest, dass es im Korpus Basen gibt, um die diese zwei Suffixe im Luxemburgischen konkurrieren, z. B. *realisabel* vs. *realiséierbar*. Beide Wörter sind im Kernkorpus ein Hapaxlegomenon, „…datt de lëtzebuergeschen Effort muss realisabel sinn…" (Parlament 2003–2004_26) und „…déi sech fir eng friddlech Koexistenz, mä eng realiséierbar Koexistenz…" (Parlament 2003–2004_17), Jedoch findet sich der erste Type 6 Mal und der zweite 18 Mal im Erweiterungskorpus. Somit wird deutlich, dass *realiséierbar* die häufigere Form der beiden Types ist. *-haft* verhält sich wie *-bar*; es gibt hier Adjektive, die definitiv aus dem Deutschen kommen, z. B. *mangelhaft, lückenhaft, gönnerhaft*. Auf der anderen Seite findet man das Suffix bei den luxemburgischen Basen, wie *feelerhaft* ‚fehlerhaft', *meeschterhaft* ‚meisterhaft', *eeschthaft* ‚ernsthaft', *niewelhaft* ‚nebelhaft'. Allerdings ist *-haft* nicht so produktiv wie *-bar* und erreicht auch nicht seine lexikalische Vielfalt an Basen. Dasselbe gilt auch für *-sam* in Lexemen deutschen Ursprungs z. B. *gemeinsam, gewaltsam* gegen luxemburgische Lexeme *spuersam* ‚sparsam', *opmierksam* ‚aufmerksam'.

2.3. Produktivität bei der Substantivbildung

Substantive besitzen unter den Wortklassen die meisten Wortbildungselemente. Das hängt damit zusammen, dass Substantive Dinge aus der objektiven Realität benennen und die meisten Types und damit die umfangreichste lexikalische Vielfalt in der Sprache besitzen. Das Luxemburgische ist in diesem Fall keine Ausnahme (Sirajzade 2013: 58). Das Diagramm 2 und die dazu gehörige Tabelle 2 zeigt, dass das Luxemburgische 6 bzw. 7 Suffixe kennt, die sehr häufig Substantive bilden, je nachdem, ob man *-heet* und *-(eg)keet* als ein oder zwei Suffixe betrachtet. Diese sind *-ung/-ong/-eng, -ioun, -echt, -(eg)keet, -heet, -age* und *-er/-ler*. Die zweite Gruppe, bestehend aus den Suffixen *-el, -ement, -nes/-nis, -enz, -ment, -teur, -esch, -in,* und *-tur/-dur,* bezieht, was die Häufigkeit betrifft, eine mittlere Position. Die Suffixe der dritten Gruppe *-mus, -schaft, -eur, -(el)chen, -(er)ei, -tum/-tem, -ist, -ert* und *-trice* kommen in der luxemburgischen Sprache seltener vor. Die vierte Gruppe besteht aus den sehr seltenen Suffixen *-(e)s, -wierk, -wiesen, -elt, -inne* und *-ling*, die im Kernkorpus in einer Anzahl unter 10 Tokens vorkommen.

Abb. 2: Tokenanzahl der substantivbildenden Suffixe im Luxemburgischen.



Das Suffix *-ung* scheint im Luxemburgischen mit Abstand das geläufigste Suffix für die Substantivbildung zu sein. Sowohl die realisierte Produktivität von 1 068/1 000 000 ≈ 0,00107 als auch die expandierende Produktivität von 491/16372 ≈ 0,029 ist am höchsten unter allen substantivbildenden Suffixen. Gefolgt wird dieses Suffix von *-ioun*, das zwar unter den substantivbildenden Suffixen nach *-ung* mit 1 404 und 407 die höchste Anzahl von Tokens und Types hat, aber mit 82 weniger Hapaxlegomena als das Suffix *-(eg)keet* besitzt (noch weniger, wenn man *-(eg)keet* und *-heet* als die Formen eines Suffixes sehen würde). Daher beträgt die realisierte Produktivität von *-ioun* 407/1 000 000 ≈ 0,000407 und die expandierende Produktivität 82/16372 ≈ 0,005. Somit hat *-ung*, statistisch gesehen keine ernsten Konkurrenten unter den substantivbildenden Suffixen in der luxemburgischen Sprache. Das Suffix *-ioun* konkurriert vielmehr mit dem germanischen Suffix *-keet/-heet*. Alleine *-keet* besitzt eine expandierende Produktivität von 123/16372 ≈ 0,0075, also eine höhere als *-ioun*. Indes ist die realisierte Produktivität bei *-ioun* höher als bei *-keet*. Dieser Unterschied fällt natürlich kleiner aus, wenn man *-keet* und *-heet* als ein Suffix betrachtet, woraus sich eine Typeanzahl von 311 ergäbe. Trotzdem erreicht diese Zahl nicht die Anzahl der Types bei *-ioun* mit 407. Laut der Anzahl der Tokens befindet sich zwischen diesen beiden das Suffix *-echt*, allerdings mit deutlich weniger Types und Hapaxlegomena. Interessant ist bei diesem typisch luxemburgischen Suffix, dass es eine erstaunlich hohe Anzahl an Tokens hat. Diese Gruppe von Suffixen endet mit *-age* und *-er* mit einer realisierten Produktivität von jeweils 64/1 000 000 ≈ 0,00 0064

und 110/1 000 000 ≈ 0,00 011 und einer expandierenden Produktivität von jeweils 24/16 372 ≈ 0,00146 und 22/16 372 ≈ 0,00124.

Tab. 2: Substantivbildende Suffixe im Luxemburgischen.

Suffix	*POS*	*Tokenzahl*	*Typezahl*	*Hapaxlegomena*
-ung/-ong/-eng	N	5187	1068	491
-ioun	N	1404	407	82
-echt	N	1175	113	44
-(eg)keet	N	1137	222	123
-heet	N	817	89	40
-age	N	541	64	24
-er/-ler	N	354	110	22
-nes/-nis	N	236	73	38
-el	N	193	56	16
-ement	N	172	52	9
-enz	N	145	44	5
-ment	N	134	13	1
-teur	N	124	31	4
-tur/-dur	N	124	46	11
-esch	N	105	15	6
-in	N	100	36	22
-(el)chen	N	61	32	20
-mus	N	40	13	4
-schaft	N	38	11	4
-eur	N	37	16	4
-(er)ei	N	25	11	2
-tum/-tem	N	18	5	1
-ist	N	17	12	5
-ert	N	13	7	2
-trice	N	12	3	1
-(e)s	N	7	4	1
-wierk	N	4	2	0
-wiesen	N	4	3	0
-elt	N	1	1	0
-inne	N	1	1	1
-ling	N	1	1	0

Schaut man sich die lexikalische Vielfalt an, die die erste Gruppe der substantivbildenden Suffixe des Luxemburgischen haben, so sieht man, dass *-ung/-ong/-eng* im Kernkorpus von nahezu über 1 000 Verbtypes Substantive

bilden kann. Größenteils sind es luxemburgische Substantive wie *Fuerschung* ‚Forschung', *Belaaschtung* ‚Belastung', *Finanzéierung* ‚Finazierung', *Ofmaachung* ‚Abmachung', *Uweisung* ‚Anweisung', *Eenegung* ‚Einigung', *Kloerstellung* ‚Klarstellung', *Opschwong* ‚Aufschwung' etc. Es finden sich aber auch Lehnwörter aus dem Deutschen, wie *Bevölkerung, Bekämpfung* oder *Handlung*. Es scheint, dass die deutschen Lehnwörter die Produktivität des Suffixes statistisch erhöhen, denn viele von diesen Lehnwörtern können im Luxemburgischen durch andere Wörter ersetzt werden, z. B. mit Lehnwörtern aus dem Französischen. Trotzdem ändert diese Tatsache nichts an der ersten Position des Suffixes bei der Produktivität. Mit einer lexikalischen Vielfalt und Flexibilität zeichnet sich das Suffix *-ioun* aus. Dieses Suffix ist eher bei den Lexemen aus dem Lateinischen zu finden und konkurriert sowohl mit *-ung*, als auch am Rande mit *-heet/-keet*, z. B. *Realisatioun* ‚Realisation' vs. *Realiséierung* ‚Realisierung', dabei ist die erste Form häufiger im Kernkorpus zu finden als die zweite, 29 vs. 5 Tokens, *Globalisatioun* ‚Globalisierung' vs. *Globaliséirung* ‚Globalisierung', diesmal mit der häufigeren Verwendung der zweiten Form, 3 vs. 9 Tokens, und *Regularisatioun/Regulatioun* ‚Regulierung' vs. *Reguléierung* ‚Regulierung', wieder zu Gunsten der ersteren Form, also 8 vs. 3 Tokens im Kernkorpus. Die zahlentechnische Konkurrenz von *-heet/-keet* und *-ioun* bestätigt sich semantisch gesehen nicht und findet daher nur am Rande statt. Beide Suffixe scheinen ihre eigenen Nischen zu haben und konkurrieren daher bezüglich der Basen nur innerhalb des Erweiterungskorpus in drei Fällen: *Absolutheet* ‚Absolutheit' – *Absolutioun* ‚Absolution', *Korrektheet* ‚Korrektheit' – *Korrektioun* ‚Korrektur', *Resolutheet* ‚Resolutheit' – *Resolutioun* ‚Beschluss'. Hierbei handelt es sich eher um eine formale Konkurrenz und der Unterschied in der Semantik der Wörter lässt es zu, dass diese Formen parallel existieren können, z. B. ist *Korrektheet* ein (idealer) Zustand im Sinne von „fehlerfrei", *Korrektioun* ist der Vorgang, der dazu führt.

Eine besondere Rolle spielt in der luxemburgischen Substantivbildung das Suffix *-echt*. Im Kernkorpus findet sich dieses Suffix in 113 Types, davon sind 44 Hapaxlegomena. Allerdings ist ein Großteil der Lexeme hier nicht dem Suffix selber zu verdanken, sondern der Tatsache, dass das Wort *Aarbecht* ‚Arbeit' im Luxemburgischen sehr kompositionsfreudig ist. *Schweessaarbecht* ‚Schweißarbeit', *Stroossenaarbecht* ‚Straßenarbeit', *Ënnerhaltsaarbecht* ‚Unterhaltsarbeit', *Botzaarbecht* ‚Putzarbeit', *Interimsaarbecht* ‚Interimsarbeit', *Gesamtaarbecht* ‚Gesamtarbeit' etc. Darüber hinaus gibt es tatsächlich Substantive im Luxemburgischen, die mit diesem Suffix gebildet werden, z. B. *Wourecht* ‚Wahrheit', *Klorecht* ‚Klarheit', *Heemecht* ‚Heimat', *Deierecht* ‚Teuerung, Preisanstieg', *Gewunnecht* ‚Gewohnheit' (auch Komposita wie *Iessgewunnecht* ‚Essgewohnheit'). Hier ist eine klare Paradigmabildung anhand der bestehenden Analogie zu sehen. *-er* ist ein klassisch luxemburgisches

Suffix zur Bildung von Substantiven aus Verben mit mehreren Funktionsklassen, auf die hier nicht weiter eingegangen wird.

Ein interessanter Fall im Luxemburgischen ist die Konkurrenz von *-esch* und *-in* zur Bildung von Feminina aus Maskulina, besonders aus Wörtern, meist Verben, die auch eine Verbindung mit *-er* eingehen können. Das erstere als substativbildendes Suffix kommt im Kernkorpus 105 Mal in 15 Types vor, davon sind 6 Hapaxlegomena. Dicht gefolgt wird dieses Suffix von *-in,* 100 Mal mit 36 Types und 22 Hapaxlegomena. Laut der Anzahl der Tokens haben die Suffixe gleiche Häufigkeit, allerdings legt letztere Gruppe mit doppelt so vielen Types und fast viermal mehr Hapaxlegomena eine höhere Produktivität an den Tag: 15/1 000 000 ≈ 0,000 015 vs. 36/1 000 000 ≈ 0,000 036 bei der realisierten Produktivität und 6/16 372 ≈ 0,000 36 vs. 22/16 372 ≈ 0,0013. Schaut man sich *-in* jedoch näher an, so findet man viele Wörter aus dem Deutschen, z. B. *Sekretärin, Beamtin, Schülerin*, aber auch viele luxemburgische Wörter, die das Suffix besitzen, wie z. B. *Kolleegin* ‚Kollegin', *Buergermeeschterin* ‚Bürgermeisterin', *Wielerin* ‚Wählerin'. Die Konkurrenz der Suffixe um die Basen zeigt sich relativ deutlich in Beispielen wie *Meeschtesch* ‚Meisterin' vs. *Meeschterin* ‚Meisterin', *Riednesch* ‚Rednerin' vs. *Riednerin* ‚Rednerin' oder *Ministesch* ‚Ministerin' vs. *Ministerin* ‚Ministerin'. Es gibt aber auch Basen, die nur mit einem der beiden Suffixe eine Verbindung eingehen, z. B. *Wielerin* ‚Wählerin', *Lëtzebuergerin* ‚Luxemburgerin', *Nopesch* ‚Nachbarin'. Bei den ersten zwei Wörtern besteht eine höhere Gefahr der Verwechslung mit dem adjektivbildenden Suffix *-esch*.

3. Technische Realisierung

3.1. Infrastruktur

Da im Institut für die Luxemburgische Sprache und Literatur bereits eine Sammlung von elektronischen Fließtexten in der luxemburgischen Sprache vorlag, war es naheliegend, die Organisation der Daten in XML zu halten. Es wurde hier absichtlich auf eine relationale Datenbank verzichtet. Stattdessen wurde Text Encoding Initiative[5] (TEI) verwendet, wobei ein wichtiger Bestandteil der Arbeit bei der Überführung der Textdaten in XML die Erstellung von TEI-Headern war. TEI stellt Richtlinien für die Strukturierung von XML-Dateien und für die Benennung der XML-Elemente zur Verfügung, um Texte, die in den Geisteswissenschaften als Forschungsobjekte dienen, zu strukturieren und zu kodieren. Der Vorteil der Verwendung des TEI liegt einerseits in der Tatsache, dass für viele Textsorten jeweils ein bereits gut überlegtes Schema verwendet werden kann. Andererseits sind auf eine solche

5 http://www.tei-c.org/index.xml, zuletzt geprüft am 29.11.2017

Art und Weise kodierte Daten nicht nur leichter zu lesen für Forscher, die diesen Standard bereits kennen, vor allem dadurch, dass TEI englischsprachige Bezeichnungen für XML-Elemente verwendet, sondern können auch bereits dafür programmierte fertige Anwendungen mit diesen Daten umgehen. Im Untersuchungskorpus wurden zu einzelnen Textdateien Metadaten erstellt, um eben diese internationale Transparenz zu gewährleisten. Die Informationen aus dem Header wurden auch in die Arbeit von anderen Skripten und Programmen integriert, sodass in jedem Schritt der Verarbeitung genau ersichtlich war, aus welchen Daten die gefundenen Belege kommen. In den Metadaten wurden z. B. Informationen bezüglich des Titels, Autors, Pseudonyms und Veröffentlichungsdatums etc. festgehalten. Folgender Auszug aus dem Korpus ist ein Beispiel für die Metadaten:

```
<TEI>
<teiHeader xml:lang="deu">
<fileDesc>
<titleStmt>
<title type="main">D'Kerfegsblo'm</title>
<title type="sub">Én Geschicht a'us dem ale Letzeborger Volleksliewen an der Muselsprôch</title>
<title type="short">D'Kerfegsblo'm 1</title>
<author> <forename>Adolf</forename> <nameLink/> <surname>Berens</surname>
<addName type="pseudonym">Ale Mann</addName> </author>
<editor> <forename/> <surname/> </editor>
</titleStmt>
<publicationStmt>
<publisher/>
<pubPlace/>
<date>1921</date>
</publicationStmt>
<sourceDesc>
<bibl>Berens, Adolf: D'Kerfegsblo'm. Én Geschicht a'us dem ale Letzeborger Volleksliewen an der Muselsprôch. vun ém ale Mann. Grevenmacher 1921.</bibl>
<file>Berens_D'Kerfegsblom 1.pdf</file>
</sourceDesc>
</fileDesc>
```

Für die Speicherung der XML-Daten wurde die XML-Datenbank eXist-db (Version 2.2) eingesetzt[6]. Eine XML-Datenbank erlaubt nicht nur das zentrale Speichern von XML-Daten und einen Zugriff über das Internet, sondern auch eine elegante Möglichkeit zur Abfrage der Daten mittels X-Query, eines W3-Standards für die Abfrage von XML-Daten (Meier 2003, S. 171). Als

6 http://exist-db.org/, zuletzt geprüft am 29.11.2017

eine native XML-Datenbank stellt eXist-db darüber hinaus zusätzliche Tools zur Verfügung, beispielsweise eine graphische Oberfläche für Browser und/oder ein eigenständiges Java-Admin-Tool, um die XML-Daten übers Netz anzuzeigen und abzufragen. Eine andere Alternative zu eXist-db stellt die XML-Datenbank BaseX dar, die ähnliche Funktionen zur XML-Abfrage zur Verfügung stellt[7]. XML-Datenbanken genießen als NoSQL-Datenbanken den Vorteil, beliebig erweiterbar zu sein, und müssen keiner rigiden einheitlichen Struktur folgen. Allerdings brauchen sie mehr Platz zur Speicherung der Daten, im Falle von eXist-db 2,7 mal mehr Speicherplatz als es beispielsweise bei MySQL der Fall ist (Freire, Sergio Miranda et al. 2016). Die Reaktionszeiten sind von der Art der Abfragen und der zurückgelieferten Datenmengen abhängig, allerdings hat BaseX laut der letzten Studien die schnellste Reaktionszeit unter den XML-Datenbanken (Freire, Sergio Miranda et al. 2016). Der Grund für eine geringere Laufzeitgeschwindigkeit im Vergleich zu anderen Datenbankarten liegt in der Natur des XML-Formats – vor allem wegen der voll ausgeschriebenen öffnenden und schließenden Elementen-Syntax. Als Format scheint JSON bis zu 30% schlanker zu sein als XML, jedoch sind hier besonders große Mengen von Daten sehr schlecht menschenlesbar.

3.2. Algorithmen und Heuristiken zur automatischen Analyse der Wortbildung

3.2.1. Automatische Suche nach Wortbildungsaffixen und Probleme

Die Ambiguität von Sprachzeichen zeigt sich auch in der Wortbildung. Es gibt im Luxemburgischen viele Wortbildungselemente, die homonym oder Homografen sind. Das Morphem *-esch* beispielsweise, kann sowohl der Bildung weiblicher Substantive wie *Nop-er+-esch* > *Nopesch* ‚Nachbarin' oder *Ried-er+-esch* > *Riednesch* ‚Rednerin' als auch der Bildung von Adjektiven wie *afrikanesch* ‚afrikanisch', *technesch* ‚technisch' oder *evangelesch* ‚evangelisch' dienen. Darüber hinaus gibt es viele Zeichenketten bei der Suche nach Wortbildungselementen, die je nach Kontext eine gänzlich andere, z. B. morphologische Funktion erfüllen, oder fester Bestandteil eines Wortstammes sind. Je kleiner das gesuchte Sprachzeichen, z. B. bei Wortbildungssuffixen *-el*, *-eg*, *-er* oder *-ei*, desto höher ist die Ambiguität. Das Element *-eg* beispielsweise, bildet Adjektive wie *dreckeg* ‚dreckig', *eekeleg* ‚eklig' oder *heefeg* ‚häufig', jedoch findet sich diese Zeichenkette auch in Substantiven wie *Mëtteg* ‚Mittag' oder *Deeg* ‚Tage' und somit in anderer Funktion. Daraus lässt sich schlussfolgern, dass eine formale Suche nach Zeichenketten oder regulären Ausdrücken sehr viele falsche Kandidaten liefern kann, die gesichtet werden müssen. Einen Teil davon kann man automatisch ausschließen, wenn man in die Suche die

7 http://basex.org/, zuletzt geprüft am 29.11.2017

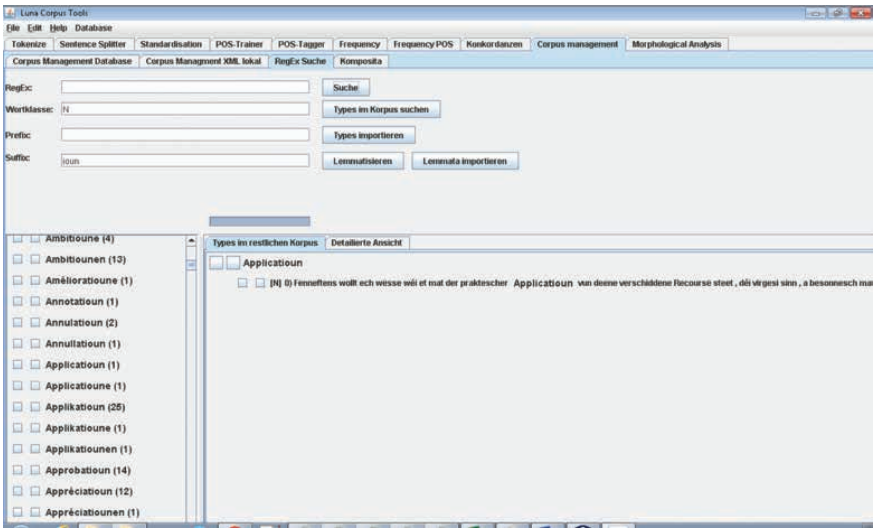
Annotation der Wortklassen miteinbezieht. Eine weitere Schwierigkeit bei der Suche nach Wortbildungssuffixen ist die Tatsache, dass diese sich sowohl im Luxemburgischen, wie auch in anderen Sprachen morphologisch gesehen näher bei dem Stamm befinden und meist vor den grammatischen Flexionselementen vorkommen. Hinzu kommen im Luxemburgischen die N-Tilgung und die uneinheitliche Schreibung, die später diskutiert werden sollen. So hilft es nicht, die formale Suche mit einem Wortende einzuschränken. Im Falle der Zeichenkette *esch*, die im Gesamtkorpus als adjektivbildendes Suffix ca. 45 000 und als substantivbildendes Suffix ca. 1 200 Mal vorkommt, findet eine formale Suche nur 32 000 Wörter, die mit der Zeichenkette *-esch* enden. Denn die restlichen Belege sind flektierte Formen, bei Substantiven Pluralendung und bei Adjektiven die jeweiligen Angleichungen an den Kasus des beschriebenen Substantivs. Das Suffix *-ei* wie bei *Roserei* ‚Ärger, Zorn' oder *Molkerei* ‚Molkerei' zu finden hat sich als einfacher erwiesen als andere Suffixe. Es gibt im Kernkorpus über 92 000 Tokens, die mit dieser Zeichenkette enden. Beschränkt man die Suche auf die minimale Größe von vier Zeichen, bekommt man nur ca. 4 000, denn viele von diesen Tokens sind Funktionswörter wie *hei* ‚hier'. Von diesen 4 000 sind aber fast die Hälfte Adjektive wie *schei* ‚scheu', *trei* ‚treu', *fonkelnei* ‚funkelneu' und Adverbien wie *elei* ‚hier' und damit uninteressant für die eigentliche Suche. Der Rest sind Substantive, aber sogar diese brauchen eine weitere Überprüfung, denn unter den Funden sind auch Wörter wie *Tierkei* ‚Türkei', die Ländernamen sind und keineswegs ein Produkt der Wortbildung. Anders verhält es sich bei dem Suffix *-eg*. Über 46 000 Tokens im Gesamtkorpus enden mit dieser Zeichenkette. Berücksichtigt man alle möglichen Flexionsendungen, steigt die Zahl der Funde auf ca. 62 000, wobei davon über 12 000 Substantive sind, die aus der Suche ausgeschlossen werden können. Zusammenfassend kann gesagt werden, dass eine bloße Suche nach Zeichenketten in einem Korpus nicht immer zu gewünschten Ergebnissen führt. Über die Regulären Ausdrücke hinaus zeigt sich hier die Verwendung der vorhandenen POS-Annotationen als eine gute Strategie. Falsche Kandidaten sind überall; auch dort, wo man sie nicht erwarten würde, z. B. bei Eigennamen wie *Akbar* bei der Suche nach *-bar* in Adjektiven wie *bemierkbar* ‚bermerkbar', *ëmsetzbar* ‚umsetztbar', oder *Ensembel* ‚Gesamtheit' und *Qualitéitslabel* ‚Qualitätslabel' bei der Suche nach *-bel* in Adjektiven wie *räsonnabel* ‚vernünftig' oder *akzeptabel* ‚akzeptabel'. Diese falschen Funde können hervorragend vermieden werden, wenn in den Suchmechanismus POS-Annotationen miteinbezogen werden. Auf diese Art und Weise wird das Programm die Zeichenkette „bar" oder „bel" nur bei den Wörtern suchen, die als Adjektiv identifiziert wurden. Substantive wie *Ensembel* oder *Qualitéitslabel* können somit aus der Suche ausgeschlossen werden, obwohl sie rein formal mit der gesuchten Zeichenkette enden.

3.2.2. Lemmatisierung

Bei der korpusbasierten Arbeit mit Wortbildungsaffixen ist eine Lemmatisierung notwendig. Wie vorher im Abschnitt „Automatische Suche nach Wortbildungsaffixen" beschrieben, kommen viele Wörter mit Wortbildungsaffixen im Korpus in flektierter Form vor. Viele Korpora in der englischen, deutschen oder französischen Sprache haben heutzutage bereits Annotationen mit Lemmata zu den jeweiligen Tokens, die hervorragend für diese Aufgabe verwendet werden können (Müller 2015: 2357). Allerdings gibt es für das Luxemburgische weder ein solch annotiertes Korpus noch verfügbare Werkzeuge, um solche Annotationen zu erzeugen. Um diese Hürde zu überwinden, stünden zwei Möglichkeiten zur Verfügung: Einen Lemmatisierer für das Luxemburgische zu implementieren oder nur die Belege zu lemmatisieren, also die Tokens, die ein Wortbildungsaffix besitzen. Im vorliegenden Fall wurde sich für die zweite Variante entschieden, zumal einerseits eine Implementierung eines Lemmatisierers für das Luxemburgische eine eigenständige Aufgabe ist und sehr lange dauern würde, andererseits besteht das Risiko, dass sich Fehler in die Analyse einschleichen. Wenn man jedoch nur die Belege lemmatisiert und dafür stochastische Methoden verwenden möchte, dann steht man vor dem bekannten „sparse data problem". Also bleibt hier ein regelbasiertes, semiautomatisches und nicht trainierbares Verfahren übrig. Diese Belege müssen ohnehin gesichtet werden, sodass entschieden wurde, diesen Schritt damit zu kombinieren.

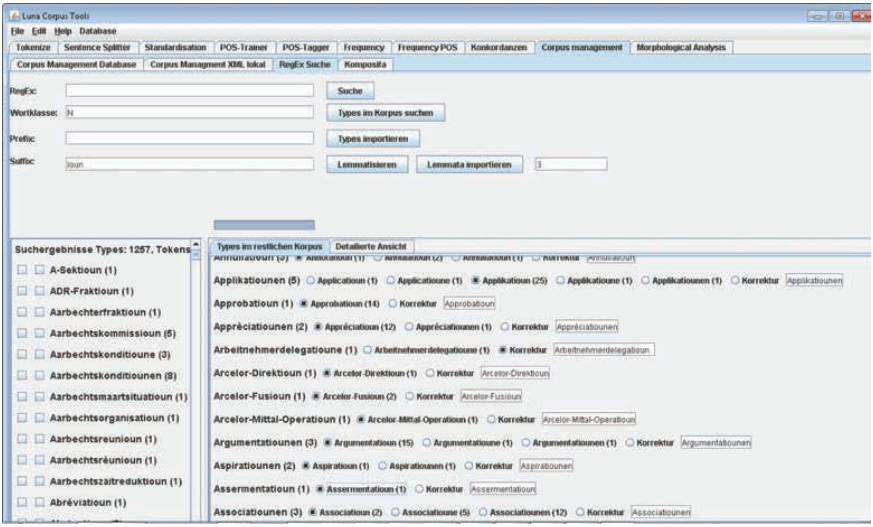
Das Verfahren, das angewendet wurde, besteht darin, dass man die Belege alphabetisch sortiert und diejenigen, die sich ähnlich sind, bündelt, ein Verfahren, das schon bei der Programmierung der ersten Lemmatisierer verwendet wurde (Klein und Rath 1971). Für die Bündelung der Belege wurde die Levenshtein-Distanz eingesetzt. Ein zusätzlicher Vorteil dieses Vorgehens besteht darin, dass man auf diese Art und Weise auch Wörter mit unterschiedlicher Schreibung bündeln kann. Wie zuvor erwähnt, besitzt das Luxemburgische mehr uneinheitliche Schreibung als die benachbarten Sprachen, obwohl schon seit 1975 eine offizielle Orthographie existiert.

Abb. 3: Überarbeitung der annotierten Belege in LuNa.



Als Beispiel für die Bündelung der flektierten Formen zu einem Lemma kann man z. B. die Types *Applicatioun*(1), *Applicatioune*(1), *Applikatioun*(25), *Applikatioune*(1) und *Applikatiounen*(1) betrachten. Hier wurde das Wort mit *c* einmal im Singular und einmal im Plural + n-Tilgung geschrieben und drei Mal mit *k*, einmal im Singular und zweimal im Plural ohne und mit n-Tilgung. In Klammern wird die Anzahl der jeweiligen Tokens in einem Type angegeben. Alle diese Tokens müssen auf ein Lemma zurückgeführt werden. Als Default bietet das Programm die häufigste Form als Lemma an, die auch mit einer Liste der möglichen Lemmata im Luxemburgischen automatisch verglichen wird. Dies muss dennoch von einem menschlichen Bearbeiter bestätigt werden, um automatisch generierte Fehler zu vermeiden. Wenn passende Lemmata nicht gefunden werden, kann der menschliche Bearbeiter das Lemma per Hand eingeben. Mit so einer Bestätigung, die meist nur einen Klick erfordert, werden in diesem Beispiel bereits 29 Wörter lemmatisiert und in der Datenbank gespeichert. Die Levenshtein-Distanz, benannt nach dem russischen Mathematiker Wladimir Iossifowitsch Lewenstein, der das Modell entwickelt hat, wird schon seit einer Weile in der linguistischen Forschung erfolgreich eingesetzt (List 2014: 145). Um die flektierten Formen eines Lemmas zu bündeln, reicht meistens eine Distanz von 3 oder 4 aus.

Abb. 4: Lemmatisierung der Belege in LuNa.



3.2.3. Basis- und Kontextanalyse

Die bisher beschriebenen Ansätze gehen meist von den Affixen aus. In Baayens Formeln werden Hapaxlegomena berücksichtigt, um vor allem die Neubildungen, die in der Sprache noch nicht häufig auftreten, in die Produktivität miteinzubeziehen. Hapaxlegomena können aber auch fossilierte Belege beinhalten, die eher das Gegenteil darstellen. Aus diesem Grund ist es sinnvoll, eine Analyse der Basen der Wortbildung durchzuführen, die man auch größtenteils automatisieren kann. Solch eine Analyse würde nicht nur bezüglich der Verwendung der Basen ohne das dazugehörige Suffix Klarheit schaffen, sondern auch helfen, die Semantik und die Funktionsklasse der Suffixe besser zu bestimmen. Die Vorgehensweise ist einfach: Nachdem die Belege identifiziert und lemmatisiert bzw. annotiert wurden, kann man automatisch die Basen aus den Lemmata ausschneiden und im Korpus nach ihrer sonstigen Verwendung suchen. Eine solche Analyse wurde von Sirajzade (2012:87) mit den Verben durchgeführt. Ob die Ergebnisse solch einer Analyse mathematisch formuliert oder gar in die Produktivitätsanalyse miteinbezogen werden können, bleibt noch zu untersuchen.

Eine Kontextanalyse ist unabdingbar, wenn man bestimmen möchte, welche Funktionen Wortbildungsaffixe haben. Wie so häufig, kann auch hier Software den Linguisten nicht komplett ersetzten. Allerdings können die Tools die linguistischen Forschungen unterstützen; einerseits um die Arbeit schneller voranzutreiben, andererseits können hier Statistiken sehr hilfreich sein, um Beobachtungen zu untermauern oder zu widerlegen. Hier kann die Software

ein Fenster aus dem Kontext der Belege schneiden und die Wörter, die darin vorkommen, nach ihrer Häufigkeit sortieren. Zur algorithmischen Kontextanalyse passt auch die berühmte Feststellung von Firth: „You shall know a word by the company it keeps!“ (Evert 2004: 15), die in vielen linguistischen Applikationen im Bereich der Semantik zitiert wurde.

4. Fazit und Ausblick

Korpuslinguistische Methoden können einen wichtigen Beitrag zur linguistischen Forschung allgemein, und so auch zur Wortbildungsforschung im Speziellen leisten. Programmatisches Vorgehen kann viele Probleme lösen: Die Suche nach Wortbildungsaffixen, die Bestimmung ihrer Häufigkeiten, ihre Verteilung in einem Korpus, Festlegung ihrer Produktivität durch grundlegende statistische Analyse, Analyse des Kontextes der Belege und der Basen im sonstigen Sprachgebrauch. Jedoch entwickeln sich diese Methoden beständig weiter. Die Entwicklungen heutzutage, besonders im Bereich der Korpora, sind gewaltig. Vorhandene Annotationen erlauben tiefergehende Suche und Analysen. Es gibt mittlerweile nicht nur viele fertige Werkzeuge, die man verwenden kann, sondern die Algorithmen und Heuristiken, auf denen diese Werkzeuge basieren, werden ständig weiterentwickelt. Bei der Speicherung der Daten empfiehlt sich XML. Damit erzeugt man zwar mehr Daten als mit z. B. JSON oder wenn man die Daten in einer MySQL-Datenbank speichern würde, jedoch sind sie besser menschenlesbar. Durch TEI verwendet man einen internationalen Standard – diese Richtlinie gibt den geisteswissenschaftlichen Daten nicht nur eine gut überlegte Struktur, sondern ermöglicht es auch, dass diese Daten international verständlich und austauschbar sind. Die genaue Dokumentation der TEI-Versionen (P5) ist außerdem eine Voraussetzung für die Nachhaltigkeit. Setzt man TEI-XML in einem Projekt mit mehreren Mitarbeitern ein, so ist es empfehlenswert eine XML-Datenbank zu verwenden, um die Daten zentral zu verwalten.

Die Wortbildung ist ein fester Bestandteil der linguistischen Forschung und dafür zuständig, dass das Lexikon einer Sprache wächst und erneuert wird. Das Luxemburgische, das sich zwischen germanischem und romanischem Sprachraum befindet, zeigt hier eine besondere Dynamik. Wie festgestellt wurde, besitzt das Luxemburgische einerseits viele Affixe, die über Lehnwörter in die Wortbildung gelangt sind, andererseits gibt es Affixe, die sehr typisch für die luxemburgische Sprache sind und die sich in benachbarten Sprachen nicht wiederfinden lassen. Bei der Bildung von Adjektiven stellen *-lech*, *-eg*, *-esch* die produktivsten Suffixe dar. Die Suffixe *-bar* und *-bel* konkurrieren um einige gemeinsame Basen. Jedoch stellt sich *-bar* statistisch gesehen als die produktivere Variante dar, denn es kann zusätzlich zu den germanischen auch mit romanischen Basen eine Verbindung eingehen. Das Umgekehrte ist bei *-bel* nicht der Fall, dieses Suffix verbindet sich meist mit Basen romanischen

Ursprungs. Die Affixe, die mit Lehnwörtern in das Luxemburgische gelangen, können die Präsenz dieser zwar stärken, aber ändern die Wortbildung prinzipiell nicht sonderlich. Das Gesagte gilt auch für die Bildung der Substantive; auch hier gibt es Affixe, die über Lehnwörter Einzug in die Sprache gehalten haben. Das romanische *-ioun* trifft auf germanische Konkurrenten in Gestalt von *-heet/-keet* und *-ung*. Während die Konkurenz mit *-heet/-keet* nur formal stattfindet und ein leichter Unterschied in der Semantik der Bildungen beiden Formen eine gleichzeitige Existenz erlaubt, stellt *-ung* einen ernsthaften Konkurrenten für *-ioun* dar und zeigt sich produktiver. Die Produktivität von *-ung* wird zusätzlich durch die Lehnwörter aus dem Deutschen gestärkt. Ein Großteil der Affixe ist jedoch als ursprünglich luxemburgisch zu bezeichnen. Sie sind moselfränkischen bzw. germanischen Ursprungs und entwickeln sich unabhängig vom benachbarten Deutschen, wie beispielsweise die Analyse der substantivbildenden Suffixe *-echt* oder *-esch* zeigt.

Literatur

Ali, Sundus Muhsin & Khalid Shakir Hussein. 2014. The Comparative Power of Type/Token and Hapax legomena/Type Ratios. A Corpus-based Study of Authorial Differentiation. *Advances in Language and Literary Studies* 5(3). 112–119.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling (ed.), *Corpus Linguistics. An International Handbook*. Volume 2 (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 29(2)). 899–919. Berlin & Boston: de Gruyter.

Belentschikow, Renate. 2015. Dictionaries. In Peter O. Müller (ed.), *Word-Formation. An International Handbook of the Languages of Europe* (Handbücher zur Sprach- und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science (HSK), 40(3)). 2333–2354. Berlin & Boston: de Gruyter.

Creutz, Mathias & Krista Lagus. 2002. Unsupervised Discovery of Morphemes. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* 6. 21–30. Stroudsburg, PA, USA: Association for Computational Linguistics (MPL '02). http://dx.doi.org/10.3115/1118647.1118650.

Dal, Georgette & Fiammetta Namer. 2015. Internet. In Peter O. Müller (ed.), *Word-Formation. An International Handbook of the Languages of Europe* (Handbücher zur Sprach- und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science (HSK), 40(3)). 2372–2386. Berlin & Boston: de Gruyter.

Evert, Stefan. 2004. *The statistics of word cooccurrences. Word pairs and collocations*. Stuttgart, Univ., Diss. http://elib.uni-stuttgart.de/opus/volltexte/2005/2371.

Freire, Sergio Miranda, Douglas Teodoro, Fang Wei-Kleiner, Eric Sundvall, Daniel Karlsson & Patrick Lambrix. 2016. Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data (PLoS ONE 11(3): e0150069). https://doi.org/10.1371/journal.pone.0150069.

Gilles, Peter. 2006. Dialektausgleich im Luxemburgischen. In Claudine Moulin & Nübling, Damaris (eds.), *Perspektiven einer linguistischen Luxemburgistik*. 1–28. Heidelberg: Universitätsverlag Winter.

Heid, Ulrich. 2015. Corpora. In Peter O. Müller (ed.), *Word-Formation. An International Handbook of the Languages of Europe* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 40(3)). 2354–2372. Berlin & Boston: de Gruyter.

Klein, Wolfgang & Rainer Rath. 1971. *Automatische Lemmatisierung. Ein Bericht*. Saarbrücken: Universität des Saarlandes.

Kohonen, Oskar, Sami Virpioja & Krista Lagus. 2010. Semi-supervised Learning of Concatenative Morphology. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. 78–86. Stroudsburg, PA, USA: Association for Computational Linguistics (SIGMORPHON '10). http://dl.acm.org/citation.cfm?id=1870478.1870488.

Lee, Young-Suk, Kishore Papineni, Salim Roukos, Ossama Emam & Hany Hassan. 2003. Language Model Based Arabic Word Segmentation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* 1. 399–406. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL '03). http://dx.doi.org/10.3115/1075096.1075147.

List, Johann-Mattis. 2014. *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press. http://sequencecomparison.github.io/.

Lulling, Jérôme. 2002. *La créativité lexicale dans la langue luxembourgeoise*. Montpellier: Th. doctorat. Etudes germaniques.

Meier, Wolfgang. 2003: eXist. An Open Source Native XML Database. In Akmal B. Chaudhri, Mario Jeckle, Erhard Rahm & Rainer Unland (eds.), *Web, Web-Services, and Database Systems:* NODe 2002 Web- and Database-Related Workshops Erfurt, Germany, October 7–10, 2002 Revised Papers. 169–183. Berlin & Heidelberg: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/3-540-36560-5_13.

Moulin, Claudine. 2004: Lëtzebuergesch, Universitét a Recherche. Lëtzebuergesch: Quo vadis? Actes du cycle de conferences. 107–119. Mamer: Melusina Conseil.

Müller, Peter O. (ed.) 2015. Word-Formation. *An International Handbook of the Languages of Europe* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 40(3)). Berlin & Boston: de Gruyter.

Pustylnikov, Olga & Karina Schneider-Wiejowski. 2010. Measuring Morphological Productivity. *Studies in Quantitative Linguistics* 5. 1–9.

Schneider-Wiejowski, Karina. 2011. *Produktivität in der deutschen Derivationsmorphologie*. Bielefeld: Universität Bielefeld Dissertation.

Sirajzade, Joshgun. 2013. *Das luxemburgischsprachige Oeuvre von Michel Rodange (1827–1876). Editionsphilologische und korpuslinguistische Analyse*. Trier: Universität Trier. http://ubt.opus.hbz-nrw.de/volltexte/2015/914/pdf/Sirajzade2013.pdf.

Straková, Jana, Milan Straka, Milan & Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 13–18. Baltimore, Maryland, Juni 2014. Association for Computational Linguistics.

Joshgun Sirajzade
University of Luxembourg
Faculty of Science, Technology and Communication
Computer Science and Communication
Maison du Nombre
6, Avenue de la Fonte
L-4364 Esch-sur-Alzette
joshgun.sirajzade@uni.lu