INSTITUT FÜR DEUTSCHE SPRACHE

Bundesministerium für Bildung und Forschung

Schlussbericht

BMBF-Forschungsprojekt
Förderkennzeichen 01UA0815

Verbundprojekt

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen

Teilvorhaben 1
Förderkennzeichen 01UA0815B

Die Basislemmaliste der neuhochdeutschen Standardsprache

Institut für Deutsche Sprache Mannheim

Projektleitung IDS: Prof. Dr. Dr.h.c. mult. Eichinger

Autorin des Schlussberichts: Heike Stadler

Verbundkoordinatorin: Prof. Dr. Andrea Rapp
Institut für Sprach- und Literaturwissenschaft, TU Darmstadt

Verbundpartner: Prof. Dr. Jörg Schulz
Theodor Boveri-Institut für Bioinformatik,
Universität Würzburg

Prof. Dr. Dietmar Seipel
Institut für Informatik, Universität Würzburg

Prof. Dr. Werner Wegstein
Institut für deutsche Philologie – Computerphilologie,
Universität Würzburg

Laufzeit des Vorhabens: 01.10.2008 – 31.12.2012

Inhaltsverzeichnis

I. Kurze Darstellung....3
1. Aufgabenstellung....3
2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde....3
3. Planung und Ablauf des Vorhabens....4
4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde....7
5. Zusammenarbeit mit anderen Stellen....8
II. Eingehende Darstellung....9
1. Erzielte Ergebnisse....9
1.1 Basislemmaliste und Eigennamenliste....9
1.1.1 Der Lemmabegriff....9
1.1.2 Konzeption der Basislemmaliste und Eigennamenliste....9
1.1.3 Sprachressourcen....13
1.1.4 Verarbeitung der Sprachressourcen – Programmmodule....17
1.2 Varietätenlemmalisten....25
1.3 Verwendete Literatur....32
2. Die wichtigsten Positionen des zahlenmäßigen Nachweises....35
3. Notwendigkeit und Angemessenheit der geleisteten Arbeit....36
4. Voraussichtlicher Nutzen....36
5. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen....37
6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses....37

Abbildungsverzeichnis

Abb. 1: BLL 1.0....11
Abb. 2: Ambige Lemmata aus der BLL 1.0 und der Eigennamenliste 1.0....12
Abb. 3: Eigennamenliste 1.0....12
Abb. 4: Linguistische Annotation des *TreeTaggers* für die Satzinitiale Wortform *Rot* (St.Galler Tagblatt 2000)....14
Abb. 5: Die Programmmodule....18
Abb. 6: Frequenzsortierte NLP-Tool spezifische „Ur“-Lemmalisten....20
Abb. 7: Frequenzsortierte NLP-Tool spezifische Lemmalisten nach Durchlaufen des Moduls 2....21
Abb. 8 Programmmodul 3....22

Tabellenverzeichnis

Tab. 1: Verteilung des Relativen Identitätswertes auf Zeitungen aus Deutschland, Österreich und der Schweiz....27
Tab. 2: Matrix mit den Frequenzen der nominalen Lemmata der einzelnen Zeitungen zur Berechnung der Euklidischen Distanz....28
Tab. 3: Lemmalistenvergleich Österreich - Deutschland 1999....30

I. Kurze Darstellung

1. Aufgabenstellung

Die Abbildung und Modellierung von Varianz wird im Projekt *Wechselwirkungen zwischen linguistischen Verfahren, Methoden und Algorithmen* auf der sprachlichen Seite u.a. repräsentiert durch die Metalemmaliste, die Lemmata der neuhochdeutschen Standardsprache mit diachronisch und diatopisch markierten Lemmata verknüpft. Die zeitlich und regional markierten Varianten stammen aus Wörterbüchern des Trierer Wörterbuchnetzes. Die Lemmata der nhd. Standardsprache werden in einer korpusgenerierten Basislemmaliste (BLL) zur Verfügung gestellt, in der neben den Lemmata auch Angaben zu deren Wortart(en) und Gebrauchshäufigkeit verzeichnet sind.

Die Lemmata der BLL bilden das Gemeinsame Dritte, auf das die Lemmata der Varietäten-Wörterbücher in der Metalemmaliste abgebildet sind, die Lemmata der BLL der nhd. Standardsprache konstituieren die Metalemmata der Metalemmaliste. Die BLL soll in ihrer Funktion als Tertium Comparationis den Sprachgebrauch im heutigen Standarddeutsch widerspiegeln. Dadurch wird sichergestellt, dass die verschiedenen Instanzen der Varietätenlemmata auf Lemmata abgebildet werden, die momentan in der Standardsprache gebräuchlich sind. Über das Metalemma lassen sich die äquivalenten Ausdrücke in den Varietäten finden, ohne dass man von deren regionalen oder historischen Ausprägungen Kenntnisse besitzt.

Die Umsetzung der semasiologischen Zugriffsmöglichkeit auf sämtliche Varietätenlemmata über ein Lemma der nhd. Standardsprache erfolgt auf der Grundlage einer XML-basierten Datenbank nach aktuellen Standards der Kodierung von Lexikoneinträgen (TEI P5). Die Metalemmaliste ist dynamisch und netzartig konzipiert, so dass immer neue Teilbereiche, Verzweigungen und Ontologien angedockt werden können (vgl. TV 2). Die Anknüpfung der Varietätenlemmata an die Lemmata der nhd. Standardsprache aus der BLL erfolgt mit Hilfe von Algorithmen, die im TV 3.2. (Informatik Würzburg) implementiert wurden.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Institut für Deutsche Sprache Mannheim, das die Aufgabe hat, die deutsche Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte wissenschaftlich zu erforschen und zu dokumentieren, ist die ideale Institution, die Basislemmaliste des Neuhochdeutschen

einzubringen. Neben dem Deutschen Referenzkorpus (DeReKo[1]) und verschiedenen Wort- bzw. Lemmalisten (DeReWo[2]) verfügt das IDS auch über ein am Institut entwickeltes umfangreiches Instrumentarium von innovativen korpusanalytischen, Struktur entdeckenden mathematisch-statistischen Methoden für deren Erschließung. Die in Textkorpora aufgezeichneten Resultate von Kommunikationsprozessen werden von der synchronen Sprachwissenschaft als empirische Grundlage und als unmittelbarer Gegenstand explorativer Forschungsarbeiten betrachtet, analog zur Sicherung und Vergegenwärtigung des kulturellen Erbes in Materialsammlungen und in Material erschließenden Grundlagenwerken durch die historischen Wissenschaften.

Durch die umfassende elektronische Sicherung seines primären Forschungsgegenstandes öffnete sich das Institut einer systematischen und tief greifenden Anwendung naturwissenschaftlicher Methoden und Technologien vor allem aus den Bereichen Mathematik, Statistik und Informatik. Seit Jahrzehnten bedient sich das IDS solcher Methoden einerseits als unverzichtbarer Arbeitsinstrumente und macht diese andererseits zum sekundären Gegenstand seiner eigenen interdisziplinären korpuslinguistischen Forschung, insbesondere im Bereich der Kookkurrenzanalysen, emergenter Grammatiken, lexikalischer Semantik oder multidimensionaler Korpusanalysen.

3. Planung und Ablauf des Vorhabens

Der Schwerpunkt des Teilvorhabens lag in der Konzeption und Umsetzung der maschinellen Erstellung einer BLL der nhd. Standardsprache aus großen digitalen Korpora. Der Umfang der BLL soll im Vergleich zur DeReWo-Liste 2007 (30.000 Einträge), die den Ausgangspunkt bildete, in einem Maße zunehmen, das garantiert, dass die Lemmata der Varietäten-Wörterbücher mit 23.000 bis 270.000 Einträgen ein Pendant in der BLL finden. Für ausgestorbene Wörter oder Dialektwörter ohne Äquivalent in der BLL werden Pseudolemmata gebildet.

Zu Projektbeginn lag erstmalig die parallele Auszeichnung von DeReKo mit linguistischen Annotationen durch mehrere Lemmatisierer, Part-of-Speech-Tagger (POS-Tagger) und Parser vor. DeReKo mit derzeit mehr als 5 Milliarden Wortformen fungiert als textuelles Fundament der BLL. Das mehrfach linguistisch annotierte Korpus ist bis dato nur IDS-intern vollständig zugänglich. Jedes der linguistischen Annotationswerkzeuge bestimmt zunächst die Wortformen im Korpus und weist jeder Wortform im Korpus mindestens ein Lemma und einen Part-of-

1 DeReKo: www.ids-mannheim.de/kl/projekte/korpora/, Kupietz/Belica/Keibel/Witt (2010).
2 DeReWo: www.ids-mannheim.de/kl/projekte/methoden/derewo.html

Speech-Tag (POS-Tag) zu. Darüber hinaus können die linguistischen Annotationen ausführliche morphologische, morphosyntaktische und syntaktische Informationen beinhalten, die Granularität und der Umfang der linguistischen Informationen variiert mit den Aufgaben und der Funktionsweise der linguistischen Annotationswerkzeuge. Bei der Erstellung der BLL sind nur das an die Wortform vergebene Lemma und der POS-Tag relevant.

Die Lemmata wurden zusammen mit ihren POS-Tags für jedes linguistisches Annotationswerkzeug (NLP-Tool – Natural-Language-Processing-Tool) separat aus dem Korpus extrahiert und in Frequenzlisten summiert. Das Format der BLL, in dem Lemma und POS-Tag den Identifikationsschlüssel bilden, um in ihrer Wortart ambige Lemmata zu unterscheiden, wird auch auf die NLP-Tool spezifischen Lemmalisten angewandt. Ergänzt werden die Lemma-POS-Einträge der NLP-Tool spezifischen Lemmalisten durch die absoluten Frequenzen ihres Auftretens im Korpus.

In die Konzeption und Programmierung der Lemmatisierer, Tagger und Parser sind unterschiedliche linguistische Theorien und sprachliche Interpretationen eingeflossen. Die aus den annotierten Korpora extrahierten Lemmalisten variieren je nach verwendetem NLP-Tool beträchtlich in Bezug auf die ermittelten Lemmata, deren POS-Tags und Frequenzen. Dies zeigt sich beispielsweise in der stark abweichenden Anzahl der Lemmata in den NLP-Tool spezifischen Lemmalisten, denen das gleiche Korpus mit separaten linguistischen Annotationen unterschiedlicher NLP-Tools zugrunde liegt. Die Anzahl der Lemmata pro NLP-Tool spezifischer Lemmaliste variiert zwischen 760.000 und 28.000.000. Auch die Tagsets sind different, es werden zwischen 12 und 54 Part-of-Speech Kategorien verwandt, um die Wortarten der Lemmata abzubilden.

In Bezug auf die Metalemmaliste ist insbesondere relevant, dass in der BLL weder gebräuchliche Lemmata fehlen, noch dass sie lexikografisch gesehen überflüssiges Sprachmaterial (Substantivierungen und Partizipien ohne Bedeutungsunterscheidung zur Grundform), die summierten Orthografiefehler der Korpora (*schießlich*), fremdsprachliche Lemmata (*the*), unbekannte Buchstabenkombinationen (*zsak*) oder Lemmatisierungsfehler (*Flitterwoche*) enthält. Durch einen einfachen Abgleich der NLP-Tool spezifischen Lemmalisten untereinander würde eine Vielzahl von lexikografisch gesehen überflüssigen oder falschen Lemmata in die BLL einfließen. Nicht nur Partizipien und Substantivierungen, sondern auch Wortformen, denen falsche Lemmata zugewiesen werden, kommen mitunter sehr häufig im Korpus vor und wären auch im vorderen Bereich der Basislemmaliste zu finden. Der

automatische Abgleich der Lemmalisten mit einer lexikografisch validierten Lemmastrecke garantiert, dass die Basislemmaliste mit den lexikografischen Konventionen korreliert.

Als lexikografisch validierte Lemmastrecke wurde dem Wechselwirkungs-Projekt die Stichwortliste von *elexiko*[3] zur Verfügung gestellt, die neben Lemmata Angaben zu deren orthografischen und morphologische Varianten enthält. *Elexiko* entstand ursprünglich selbst aus einer korpusbasierten Stichwortkandidatenliste, die durch eine „redaktionelle lexikologisch-lexikografischen Überprüfung und Aufbereitung“ (Schnörch 2005: 75) in die *elexiko* Stichwortliste überführt wurde. Durch den Vergleich mit Lemmastrecken weiterer Wörterbüchern wurde sichergestellt, dass keine wichtigen Lemmata in *elexiko* fehlen. Die jahrelange fundierte lexikografische Arbeit an *elexiko* wurde im Wechselwirkungs-Projekt genutzt, um den Fokus der manuellen Nacharbeiten auf die Verifizierung bzw. Abweisung von Lemmakandidaten mit einer hohen Frequenz zu legen, die noch nicht in *elexiko* enthalten sind. Durch das sukzessive Einfließen neuer aus dem Korpus gewonnener Lemmata in die Lemmastrecke von *elexiko* entstand die Wechselwirkungs-Stichwortliste, die ca. 25.000 zusätzliche Lemmata enthält.

Mit der mehrfachen linguistischen Annotation von DeReKo und der digital verfügbaren Stichwortliste von *elexiko* besteht erstmalig die Möglichkeit, eine korpusbasierte, frequenzsortierte Basislemmaliste der nhd. Standardsprache mit POS-Angaben maschinell zu erstellen. Den verifizierten Lemmata werden automatisch die POS-Tags zugewiesen, die bei einer vorher definierten Anzahl der NLP-Tools übereinstimmen. Für die verbleibenden Zweifelsfälle wird über die Vergabe von Konfidenzwerten an die NLP-Tools für das POS-Tagging sichergestellt, dass der POS-Tag von den NLP-Tools mit einem hohen Konfidenzwert in die BLL übernommen wird.

Das am IDS konzipierte Modell zur Erstellung der BLL der nhd. Standardsprache zur Verarbeitung der Sprachressourcen besteht aus Programmmodulen zur Analyse, Vereinheitlichung und Komparation der linguistischen Informationen. Es löst die weitgehend automatisierte Generierung der Basislemmaliste aus annotierten Korpora parametrisierbar und modular. Das Modell umfasst die Extraktion der Lemmata und der zugehörigen POS-Tags aus den annotierten Korpora, den Abgleich der Tagsets der verwendeten linguistischen Annotationswerkzeuge, den Vergleich der NLP-Tool spezifischen Lemmalisten mit der unter lexikografischen Gesichtspunkten zusammengestellten Wechselwirkungs-Stichwortliste, das Alignment unterschiedlicher Nennformen eines Lemmas zwischen den NLP-Tool spezifischen

3 *elexiko*: www.owid.de/wb/elexiko/start.html, Haß (2005), Klosa (2011).

Lemmalisten, die wahlweise Subsumption von Partizipien und Substantivierungen unter der entsprechenden Verbform, die Rückführung orthografischer Varianten auf ein Lemma und die Bereitstellung von Sonderlemmalisten. Das Korpus sowie die linguistischen Annotationswerkzeuge sind austauschbar. Durch die Wahl von DeReKo als Korpus wird garantiert, dass die BLL den allgemeinen öffentlichen Gebrauch der schriftlichen Standardsprache repräsentiert. Die im Projektzeitraum entstandene BLL der nhd. Standardsprache mit 325.000 Einträgen ist nach der Häufigkeit der Lemmata sortiert und enthält Informationen zur Wortart der einzelnen Lemmata. Eine Eigennamenliste mit 68.000 Einträgen wird zusätzlich zur Verfügung gestellt.

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Die für das Verbundprojekt relevante, im Jahr 1999 im Programmbereich Korpuslinguistik des IDS (halb)automatisch aus dem DeReKo-Archiv generierte Version der DeReWo-Referenzlemmaliste umfasste rund 300.000 Lemmata der neuhochdeutschen Standardsprache und liegt auch der Lemmaliste des lexikalisch-lexikologischen korpusbasierten Informationssystems des IDS (*elexiko*) zugrunde. Im Jahr 2007 wurden am IDS Arbeiten an einer Open Source Neuauflage der DeReWo-Liste aufgenommen. Diese umfasst die 30.000 häufigsten (korpusbasiert ermittelt) deutschen Lemmata und ist der definierte erste Kern der Metalemmaliste des Verbundvorhabens. Das IDS bettet seine Mitwirkung am Wechselwirkungs-Projekt organisatorisch und methodisch in den Programmbereich Korpuslinguistik und dessen aktuellen Forschungskontext ein.

Strategien zur Ermittlung, Analyse und lexikografischen Aufbereitung des zentralen Deutschen Wortschatzes mit ausführlichen Einblicken in die Konzeption von Grundwortschätzen, Häufigkeitswörterbüchern oder Frequenzwortlisten werden in Schnörch (2002) beschrieben. Gemeinsam ist den vorgestellten Lemmalisten die Korpusbasiertheit und die händische Lemmatisierung und Wortartenbestimmung der Wortformen. In dem 2006 erschienen Frequenzwörterbuch des deutschen Kernwortschatzes für den Fremdspracherwerb von Randall/Tschirner bildet eine automatische Lemmatisierung und Wortartenbestimmung die Grundlage, neben *TreeTagger* wird das *Wordsmith-Tool* Version 3 angewandt. Trotzdem wird die Disambiguierung von Wortformen und die Lemmatisierung größtenteils per Hand erstellt (Randall/Tschirner 2006: 139). Der Korpusumfang wird auf 4,2 Mio. Wortformen festgelegt, damit das Korpus „noch per Hand bearbeitet und werden konnte und dessen Wortlisten lemmatisiert werden konnten“ (Randall/Tschirner 2006: 137).

Die Anzahl der von Randall und Tschirner (2006) aufgeführten Lemmata beträgt etwas über 4.000, als Aufnahmekriterium wird das 16-malige Vorkommen eines Lemmas im Korpus verwandt. Die in Schnörch (2002) erwähnten Lemmalisten haben einen Umfang zwischen 1.000 und 16.000 Einträgen. Die 2009 erschienene auf DeReKo basierende Wortgrundformliste in der Reihe DeReWo des IDS Mannheim verzeichnet 40.000 Einträge ohne Wortartenangaben. Für die automatische Lemmatisierung von DeReKo wurde hier der Lemmatisierer *glemm* eingesetzt. Die lexikografischen Kriterien für die händische Bearbeitung der automatisch erstellten Lemmaliste zur Überführung in die Wortgrundformliste des Deutschen werden in der Benutzerdokumentation zu DeReWo dargestellt.

5. Zusammenarbeit mit anderen Stellen

Die technologische Abstimmung mit den Projektpartnern erfolgte kontinuierlich über den gesamten Zeitraum während der regelmäßig stattfindenden JourFixe und Arbeitstreffen. Innerhalb des Verbundprojektes wurden verschiedene Dienste und Sonderlemmalisten zur Verfügung gestellt. Die Abbildung der Lemmata der Varietäten-Wörterbücher auf die Lemmata der nhd. Standardsprache, ist für deren Verknüpfung in der Metalemmaliste elementar. Sie funktioniert wesentlich über die morphematische Zerlegung der Lemmata der Basislemmaliste, die morphologische Segmentierung der Lemmata der Varietäten-Wörterbücher und das Alignment zwischen den die Lemmata konstituierenden Morphemen. Vor der Auswahl des geeigneten Tools für die Morphemzerlegung von Varietäten-Lemmata und nhd. Standard-Lemmata stellte das IDS in Abstimmung mit TV2 und TV3.2 hausinterne morphologische Analysetools als Vergleichs- und Arbeitsgrundlage zur Verfügung. Für die Zuordnung varianter Lemmata spielen des Weiteren einfache Stringvergleiche, die Anwendung von Lautgesetzen, bereits in den Wörterbüchern bestehende Verweise und Zeichenersetzungen eine Rolle.

TV 3.1 behandelt einen weiteren Aspekt der Klassifizierung von Varianz und Basis, der den Zusammenhang zwischen der Kombinationsfähigkeit von Morphemen und der Gebrauchshäufigkeit von Wörtern erfasst. Die für das Teilvorhaben 3.1 bereitgestellten Test- und Sonderlemmalisten mit den absoluten Frequenzen aus den Korpusdaten dienen u.a. als Grundlage für die Bestimmung der Korrelation zwischen Vernetzungsgrad von Morphemen und Wortgebrauch. Die morphematische Segmentierung verschiedener Sonderlemmalisten ermöglicht exemplarische Netzwerkvergleiche zwischen biologischen und linguistischen Daten.

II. Eingehende Darstellung

1. Erzielte Ergebnisse

1.1 Basislemmaliste und Eigennamenliste[4]

1.1.1 Der Lemmabegriff

Als 'Lemma' wird in der Lexikografie das Stichwort in einem Nachschlagewerk bezeichnet. Die Zitierform im Wörterbuch ist üblicherweise identisch mit der Wortgrundform, auf die die flektierten Formen eines Wortes zurückgeführt werden. Für Nomen entspricht die Grundform im Deutschen dem Nominativ Singular, für Verben dem Infinitiv und für Adjektive dem Positiv. Die Reduktion der Flexionsformen eines Wortes auf die Wortgrundform wird 'Lemmatisierung' genannt. Die Lemmata der Stichwortliste im Wörterbuch werden durch 'Lemmaselektion' ermittelt. Im Folgenden werden als 'Lemmata' sowohl die Grundformen bezeichnet, die den Wortformen im Korpus durch ein linguistisches Annotationswerkzeug zugeordnet werden (Stichwortkandidaten), als auch die in einem Wörterbuch oder einer Wortgrundformenliste verzeichneten Stichwörter.

Lemmatisierungskonzepte aus lexikografischer Sicht sind in den folgenden Artikeln eingehend erläutert: Gallmann (1991), Knowles/Zuraidah (2004), Rosengren (1969), Schnörch (2005), Wiegand (1983). Die maschinelle Umsetzung der linguistischen Vorgaben zur Lemmatisierung zeigt jedoch ein mannigfaltiges Spektrum an Interpretations- und Fehlermöglichkeiten. Zur Erstellung einer unter lexikografischen Gesichtspunkten validen korpusbasierten Wortgrundformenliste ist ein Abgleich der automatisch aus den linguistisch annotierten Korpora extrahierten Lemmalisten mit der digital verfügbaren Lemmastrecke eines Wörterbuches unumgänglich.

1.1.2 Konzeption der Basislemmaliste und Eigennamenliste

Mit der linguistischen Annotation von DeReKo[5] durch mehrere NLP-Tools und der Verfügbarkeit der Stichwortliste von *elexiko*[6], dem Online-Wörterbuch der Deutschen Gegenwartssprache, besteht die Möglichkeit, eine frequenzsortierte Basislemmaliste mit Wortartenabgaben der nhd. Standardsprache weitgehend maschinell zu generieren. Mit mehr

4 Teile der eingehenden Darstellung der erzielten Ergebnisse sind Stadler (2013a) entnommen. In der Publikation werden zusätzliche Informationen gegeben, zahlreiche Beispiele für die automatische Lemmatisierung und das POS-Tagging linguistischer Annotationswerkzeug genannt und weitere Aspekte der Erstellung der BLL behandelt.

5 DeReKo: www.ids-mannheim.de/kl/projekte/korpora/, Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010).

6 *elexiko*: www.owid.de/wb/elexiko/start.html, Haß (2005), Klosa (2011).

als 325.000 Lemmata besitzt die BLL der nhd. Standardsprache Version 1.0 (2012) deutlich mehr Einträge als bisherige Häufigkeitswörterbücher und Wortgrundformenlisten für das Deutsche. Die automatische Lemmatisierung der Wortformen und das POS-Tagging ermöglichen die Bearbeitung großer Textkorpora. Die automatisch generierten Lemmata und POS-Tags korrelieren jedoch nur bedingt mit dem lexikografischen Standard, die Lemmatisierungskonzepte und Lemmaformen der linguistischen Annotationswerkzeuge unterscheiden sich von den üblichen Konventionen für die Erstellung von Wörterbüchern (vgl. Stadler 2013a, 2013b).

Die BLL stellt keine einfache Abbildung der maschinellen Lemmatisierung der Wortformen im Korpus dar, sie soll häufige und lexikografisch relevante Wörter des Deutschen beinhalten. Die automatisch generierten frequenzsortierten Lemmalisten der NLP-Tools sind zur Darstellung des standardsprachlichen deutschen Wortschatzes nicht geeignet. Sie enthalten aus dem Korpus extrahierte Zeichenkombinationen, die keine Wörter bezeichnen, Orthografiefehler und fremdsprachliches Material sowie durch die automatische Lemmatisierung produzierte Fehler und eine Vielzahl an Lemmata, die lediglich Wortformen einer Flexionsreihe anderer Lemmata darstellen und unter diesen zu subsumieren sind. Die automatisch aus dem Korpus generierten Lemmalisten entsprechen nicht der nhd. Standardsprache, eine komplette händische Nachbearbeitung der rein frequenzbasierten automatisch generierten Lemmalisten wäre daher unerlässlich.

Die lexikografische Validität der automatisch generierten Einträge wird über den Abgleich mit der Wechselwirkungs-Stichwortliste (WW-Stichwortliste) garantiert. Ungeeignete Lemmata, die nicht der nhd. Standardsprache angehören, werden auf diesem Wege aussortiert. Die WW-Stichwortliste besteht aus der Stichwortliste von *elexiko,* der Bezug zum aktuellen Sprachgebrauch wird über die händische Verifizierung zusätzlicher frequenter Lemmata aus dem Korpus hergestellt. Lemmata, die häufig im Korpus vorkommen, jedoch nicht in der Stichwortliste von *elexiko* enthalten sind, werden in die WW-Stichwortliste nach einer manuellen Überprüfung sukzessive eingegliedert (ca. 25.000). Die Kriterien für die Aufnahme von Lemmata in die WW-Stichwortliste korrespondieren im Wesentlichen mit den lexikografischen Richtlinien, die der Erstellung der Stichwortliste von *elexiko* zu Grunde liegen.[7]

7 In den beiden Sammelbänden zu elexiko von Haß (2005) und Klosa (2011) sind für die Konzeption der Stichwortliste besonders die Artikel von Schnörch (2005), Erb (2005), Klosa (2005, 2011) und Klosa/Schoolaert (2011) interessant.

Eigennamen, Abkürzungen und Buchstaben sind in der BLL nicht enthalten. Sie werden mit Hilfe der POS-Tags und regulärer Ausdrücke aus den Lemmalisten der einzelnen NLP-Tools ausgefiltert und daher auch nicht in die WW-Stichwortliste aufgenommen. Eine zusätzliche Eigennamenliste mit 68.000 Einträgen wird als Ergänzung zur BLL zur Verfügung gestellt. Die automatische Lemmatisierung und das POS-Tagging von Abkürzungen fällt zu heterogen aus, um eine automatisch generierte Abkürzungsliste zu erzeugen. Abkürzungen (Akronyme der Schriftsprache) und Buchstaben werden aus der BLL lediglich ausgegliedert. Kurzwörter (*Alu, Auto, Lkw, GPS, Kfz*) bleiben in der BLL erhalten. Sie werden von den NLP-Tools mit einem POS-Tag ausgezeichnet, der der Wortart ihrer Langformen entspricht.

Die BLL der nhd. Standardsprache ist absteigend nach der Häufigkeit der Lemmata sortiert und verzeichnet zu jedem Lemma den POS-Tag sowie die Häufigkeitsklasse der Lemma-POS-Tag Kombination.[8] In der ersten Spalte der BLL steht die fortlaufende Identifikationsnummer, es folgen das Lemma, der POS-Tag und die Häufigkeitsklasse. Die Berechnung der Häufigkeitsklassen sowie deren Vorteile gegenüber der Angabe von absoluten Frequenzen werden in der Benutzerdokumentation zu DeReWo ausführlich erläutert. Die aktuelle Version 1.0 der BLL wird unter DeReWo - dem IDS-Portal für korpusbasierte Grund- und Wortformenlisten - frei zur Verfügung gestellt.[9]

1	der,die,das	ART	0
2	in	PREP	2
3	und	CONJ	2
4	sein	V	3
5	ein(e)	ART	3
6	werden	V	3
7	von	PREP	3
8	mit	PREP	3
9	der,die,das	PRON	3
10	haben	V	3
11	im	PREP	4
...			
326934	Zetetikerin	NN	29
326935	Zökum	NN	29
326936	zollhoch	ADJ	29
326937	zwerch	ADJ	29
326938	Zwergewerfen	NN	29
326939	zwiegeschlechtig	ADJ	29
326940	Zwiegeschlechtigkeit	NN	29
326941	Zwiegeschlechtlichkeit	NN	29
326942	Zwölfflach	NN	29
326943	zyanenblau	ADJ	29
326944	Zyansäure	NN	29
326945	Zyklogramm	NN	29
326946	Zyrtolith	NN	29

Abb. 1: BLL 1.0

8 Innerhalb einer Häufigkeitsklasse ist die BLL nach der absoluten Frequenz und innerhalb einer absoluten Frequenz alphabetisch absteigend sortiert.

9 BLL 1.0: http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip

Lemma und POS-Tag bilden eine Einheit (Identifikationsschlüssel), damit ambige Lemmata eindeutig zu bestimmen sind. Unter einem ambigen Lemma wird ein Lemma mit mehreren POS-Tags verstanden. Die Auflösung von Homographen erfolgt ausschließlich über den POS-Tag. Sowohl Homographen, die ihre Verschiedenheit nur durch die Aussprache markieren (Tenor, Tenor), als auch Homographen, die eine abweichende morphologische Zerlegung aufweisen (*Stau-becken, Staub-ecken*), werden unter einem Lemma zusammengefasst. Ebenso erhalten Polyseme nur einen Eintrag in der BLL.

4	sein	V	3
23	sein	PRON	4
685	modern	ADJ	10
20619	modern	V	16
6537	Hirsch	NN	13
2127	Hirsch	NE	14

Abb. 2: Ambige Lemmata aus der BLL 1.0 und der Eigennamenliste 1.0

Zahlreiche Nomen stehen sowohl in der BLL mit dem POS-Tag NN als auch in der Eigennamenliste mit dem POS-Tag NE. Eine ambige Verwendungsweise von Nomen im Deutschen tritt durch die Gleichlautung von Gattungsbezeichnungen und Eigennamen häufig auf. Das Format für die Eigennamenliste entspricht dem Format der BLL, die Häufigkeitsklassen werden mit der Frequenz des häufigsten Lemmas der BLL berechnet, um so eine direkte Vergleichsmöglichkeit zu gewährleisten.

1	Deutschland	NE	7
2	Berlin	NE	8
3	Peter	NE	8
4	Michael	NE	9
5	Thomas	NE	9
6	Österreich	NE	9
7	München	NE	9
8	Europa	NE	9
9	Bad	NE	9
10	Wien	NE	9
11	Hans	NE	9
12	Wolfgang	NE	9
13	Frankfurt	NE	9
14	Martin	NE	9
…			
68882	Zwergengasse	NE	29
68883	Zwergenstraße	NE	29
68884	Zwerghausen	NE	29
68885	Zwiebackshausen	NE	29
68886	Zwiefaltbach	NE	29
68887	Zwingenbach	NE	29
68888	Zwinggasse	NE	29
68889	Zwirnereigasse	NE	29
68890	Zwischenposthofstraße	NE	29
68891	Zwölfschlösserhausen	NE	29
68892	Zwoschwitzbach	NE	29

Abb. 3: Eigennamenliste 1.0

Der Abgleich der automatisch generierten Lemmalisten der einzelnen NLP-Tools mit einer Stichwortliste reduziert die Frequenz jedoch auf die Sortiermöglichkeit einer Liste, deren Einträge a priori feststehen. Als Aufnahmekriterium für ein Lemma aus dem Korpus in die BLL dient nicht die Größe der Frequenz, sondern das Vorkommen eines Lemmas in der WW-Stichwortliste. Im Wechselwirkungs-Projekt bildet die BLL ein Referenzwortverzeichnis für Wörterbücher mit regional und zeitlich markierten Varietätenlemmata. Die Abbildung der nhd. Standardsprache in der BLL ist daher das primäre Ziel.

Die jahrelange fundierte lexikografische Arbeit an der *elexiko*-Stichwortliste wurde im Wechselwirkungs-Projekt genutzt, um den Fokus der manuellen Nacharbeiten auf die Verifizierung bzw. Abweisung von Lemmakandidaten aus dem Korpus mit einer hohen Frequenz zu legen, die noch nicht in der *elexiko*-Stichwortliste stehen. Die zusätzlich aus dem Korpus gewonnenen Lemmata repräsentieren in besonderem Maße die Lexik von Themen, die den Sprachgebrauch im Korpus in den vergangenen Jahr(zehnt)en prägten. In der *elexiko*-Stichwortliste sind über die Datenmodellierung in XML des Weiteren Angaben zu den orthografischen Varianten eines Lemmas in der aktuellen Standardschreibweise vorhanden. Diese Informationen werden verwendet, um orthografische Varianten im Korpus zu identifizieren und auf die heutige Standardschreibweise abzubilden. In der BLL stehen die automatisch addierten Frequenzen der Orthografievarianten bei dem Lemma, das dem Standard entspricht und das als einziges in der BLL verzeichnet ist.

Von den zahlreichen Hapax legomena und niedrigfrequenten Lemmata im Korpus stehen nur wenige in der *elexiko*-Stichwortliste. Die pauschale Aufnahme der niedrigfrequenten Lemmata der *elexiko*-Stichwortliste in die BLL wird durch ihre meist gegebene lexikografische Relevanz begründet. Während es sich bei den Lemmata, die aufgrund ihres häufigen Vorkommens zusätzlich zu den Stichwörtern aus *elexiko* in die WW-Stichwortliste aufgenommen werden, fast ausschließlich um transparente Komposita handelt (vgl.1.1.4), konstituieren die niedrigfrequenten Lemmata der BLL überwiegend Lexeme, deren Bedeutung nicht aus den einzelnen Bestandteilen erschlossen werden kann. Kenngrößen wie zeitliche Stabilität, hohe Wortbildungspotenz oder regionale Gleichverteilung eines Lemmas, die ebenfalls automatisch anhand der Korpusgrundlage ermittelt werden können, spielen in dem im Folgenden vorgestellten Ansatz keine Rolle für die Aufnahme eines Lemmas in die BLL.

1.1.3 Sprachressourcen

Bei der Erstellung der BLL werden drei unterschiedliche Arten von Sprachressourcen eingesetzt:

- Digitale Sprachkorpora
  - DeReKo
- Linguistische Annotationswerkzeuge (NLP-Tools)
  - Lemmatisierer *glemm*
  - POS-Tagger *TreeTagger, Machinese Phrase Tagger*
  - Parser *Machinese Syntax, Xerox Incremental Parser XIP*
- Digitale Stichwortlisten von Wörterbüchern
  - *elexiko*

DeReKo wurde von mehreren linguistischen Annotationswerkzeugen (NLP-Tools) parallel mit linguistischen Informationen angereichert und die Auszeichnungen in separaten Stand-Off Annotationen im XML-Format gespeichert. Jedes NLP-Tool bestimmt zunächst die Wortformen im Korpus (Tokenisierung) und weist jeder Wortform mindestens ein Lemma und einen POS-Tag zu.

```
<sentence>
    <lexeme id="6" pos="15414" len="3">
        <surface-form>Rot</surface-form>
        <sense id="0">
            <base-form>rot</base-form>
            <part-of-speech conf="0.730213">ADJD</part-of-speech>
        </sense>
        <sense id="1">
            <base-form>Rot</base-form>
            <part-of-speech conf="0.269787">NN</part-of-speech>
        </sense>
    </lexeme>
```

Abb. 4: Linguistische Annotation des *TreeTaggers* für die Satzinitiale Wortform *Rot* (St.Galler Tagblatt 2000)

Die in den annotierten Korpora enthaltenen Lemmata und ihre zugehörigen POS-Tags werden mit Perl-Skripten extrahiert und in NLP-Tool spezifischen Lemmalisten summiert. Über den automatischen Vergleich der NLP-Tool spezifischen Lemmalisten untereinander und mit der WW-Stichwortliste wird die BLL generiert (vgl.1.1.4).

DeReKo

Als Korpusgrundlage für die Erstellung der BLL dient das gesamte Deutsche Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache. Es umfasste zum Entstehungszeitpunkt der BLL Version 1.0 mehr als 5 Milliarden Wortformen.[10] Die öffentliche Schnittstelle zu DeReKo ist COSMAS-II, das *Corpus Search, Management and Analysis System* des IDS Mannheim.[11] Über COSMAS-II ist eine Auswahl aus den DeReKo-Daten abfragbar, die BLL 1.0 basiert auf dem gesamten in DeReKo enthaltenem Sprachmaterial. Die Korpusauszeichnungen in DeReKo

10 Für die Erstellung der BLL Version 1.0 wurde das DeReKo-Release 2012-I vom 29.2.2012 verwendet. Chronik der DeReKo-Freigaben: www.ids-mannheim.de/kl/projekte/korpora/releases.html
11 COSMAS-II: www.ids-mannheim.de/cosmas2/

werden unter dem Gesichtspunkt ihres Nutzens für die Lexikografie in Klosa/Kupietz/Lüngen (2012) behandelt und nach inhaltlichen Gesichtspunkten gruppiert: Annotation der hierarchischen Korpus- und Textstruktur, linguistisches Tagging, bibliographische Metadaten, textlinguistische Klassifikation, textspezifische deskriptiv-statische Metadaten und textspezifische Metadaten zu Duplikaten. Für die Erstellung der BLL spielen ausschließlich die linguistischen Annotationen eine Rolle.

DeReKo besteht aufgrund der Urheberrechte überwiegend aus Zeitungstexten aus Deutschland, Österreich und der Schweiz. Daneben sind in geringerem Umfang auch literarische Texte, Gebrauchstexte, fachsprachliche Texte und politische Texte vertreten. Ab 2012 sind die Wikipedia-Artikel und -Diskussionen in das Korpus integriert. Die Schriftsprache in DeReKo ist fast ausschließlich auf eine öffentliche, formelle Gesprächssituation beschränkt, das Korpus enthält keine Tweeds, E-Mails, Internetforen, etc. Dialektal geprägte Schriftsprache tritt nur sporadisch in Kolumnen und Zitaten der Zeitungskorpora sowie als wörtliche Rede in den literarischen Texten auf. Ausgerichtet auf Kontexte großer Öffentlichkeit repräsentiert DeReKo den Gebrauchsstandard des Neuhochdeutschen seit den 1950er Jahren, mit einem sehr deutlichen Zuwachs der Wortformen ab der Jahrtausendwende, als sich die digitale Archivierung endgültig durchsetzte. Lexikalische Besonderheiten des österreichischen und Schweizer Standarddeutsch (*Paradeiser* Ö, *grillieren* CH) werden in die BLL als selbständiges Lemma übernommen ohne Kennzeichnung als regionale Standardvariante. Orthografische Varianten (z.B. *ss* für *ß* in der Schweiz) werden unter der bundesdeutschen Standardschreibweise subsumiert.

Linguistische Annotationswerkzeuge (NLP-Tools)

Bei der Erstellung der BLL werden die Lemmata und POS-Tags verarbeitet, die den Wortformen im Korpus von unterschiedlichen linguistischen Annotationswerkzeugen automatisch zugewiesen wurden. Aus den für jedes NLP-Tool separat gespeicherten linguistischen Korpusauszeichnungen wird im ersten Schritt für jedes NLP-Tool einzeln eine Lemmaliste extrahiert, die die Lemmata mit zugehörigen POS-Tags und die akkumulierten Frequenzen für eine Lemma-POS-Tag Kombination enthält. Neben dem Lemma und dem POS-Tag können im annotierten Korpus weitere morphologische, morphosyntaktische oder syntaktische Angaben vorhanden sein, die für die Erstellung der BLL jedoch nicht ausgewertet werden.[12] Ein Bezug der Lemmata in den Lemmalisten zu Wortformen in einer bestimmten

12 Eine Ausnahme bilden Angaben zu Eigennamen und Abkürzungen. Deren Spezifikation wird bei zwei der NLP-Tools unter den *Features* gekennzeichnet. In diesem Fall werden die *Features* ebenfalls ausgewertet und

Position im Korpus ist nicht mehr herzustellen. Zusätzlich gespeichert werden die Wortformen, die das Flexionsparadigma eines Lemmas konstituieren, um Entscheidungen der NLP-Tools bei der Lemmatisierung nachzuvollziehen. Mit der Kombination aus Wortform, Lemma und POS-Tag können alle Korpusstellen gefunden werden, die den Kontext für die Lemmatisierung und Wortartenzuweisung zu einer bestimmten Wortform bieten.

Als linguistische Annotationswerkzeuge kamen ein Lemmatisierer, zwei POS-Tagger und zwei Dependenz-Parser zum Einsatz. Der Lemmatisierer *glemm* ist nur IDS-intern verfügbar, eine öffentliche Dokumentation gibt Einblick in Designprinzipien, Implementierung und linguistischen Hintergrund.[13] Von der Firma *Connexor Oy's* werden aus der Reihe *Machinese Products* zwei Softwareprodukte verwendet: *Machinese Phrase Tagger (MPT)* und *Machinese Syntax (MS)*. Die beim Kauf der Software von *Connexor* mitgelieferte Dokumentation ist nicht öffentlich zugänglich, auf dem *Cosmas II* Portal des IDS werden die Informationen aus der Dokumentation zum morphologischen und syntaktischen Tagset wiedergegeben.[14] Einen Einblick in die Arbeitsweise der *Machinese* Software von *Connexor* erhält man auf der Firmenwebseite.[15] *TreeTagger* ist ein frei verfügbarer am IMS Stuttgart entwickelter Tagger, der die Wahrscheinlichkeit der POS-Tags mittels Entscheidungsbäumen berechnet. Das Stuttgart-Tübinger Tagset und die Vorgehensweise des *TreeTaggers* sind ausführlich dokumentiert.[16] Für die kommerzielle Software *Xerox Incremental Parser* (*XIP*) gibt es Dokumentationen auf der Firmenwebseite.[17] Dort besteht auch die Möglichkeit, Beispielsätze mit dem regelbasierten Dependenz-Parser zu analysieren.[18] Bei den Parsern liegt der Schwerpunkt der linguistischen Analyse auf der Auszeichnung der syntaktischen Strukturen im Satz, die morphosyntaktische Analyse der Wortformen ist jedoch unabdingbar für die Ermittlung der grammatischen Beziehungen. Mit Lemmata und POS-Tags werden die Wortformen im Korpus von allen fünf NLP-Tools gleichermaßen ausgezeichnet.

elexiko

Die Stichwortliste des *Online-Wörterbuchs der Deutschen Gegenwartssprache – elexiko* des IDS Mannheim mit 300.000 Lemmata bildet die Grundlage für die lexikografische Validierung

der zugehörige POS-Tag konform zu den Konventionen der weiteren NLP-Tools umgewandelt zu 'NE' bzw. 'ABBR'.

13 Belica (1994).

14 Tagset Morphologie *Machinese*: www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html
Tagset Syntax *Machinese*: www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/syntax.html

15 *Machinese* Demo: www.connexor.com/nlplib/

16 *TreeTagger*: http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html, Schmid (1994), (1995), Schiller/Teufel/Thielen (1995).

17 *XIP* Dokumentation: http://open.xerox.com/Services/XIPParser/Pages/Using XIP

18 *XIP* Demo: http://open.xerox.com/Services/XIPParser/Consume/64

der Lemmata in den NLP-Tool spezifischen Lemmalisten. *elexiko* ist ab Ende der 1990er Jahre entstanden, als sich mit dem World Wide Web das Potenzial eines Online-Wörterbuchs zeigte. In den beiden Sammelbänden von Haß (2005) und Klosa (2011) wird *elexiko* ausführlich dokumentiert. Ausgangspunkt für die Auswahl der Lemmata in *elexiko* war eine Stichwortkandidatenliste, die korpusbasiert ermittelt wurde. Als Korpus dienten große Teile von DeReKo, die Wortgrundformen wurden den Wortformen vom Lemmatisierer *glemm* automatisch zugewiesen. Die Stichwortkandidatenliste enthält nur Wortgrundformen, die mindestens acht mal im Korpus vorkommen. Sie wurde durch eine redaktionelle lexikologisch-lexikografische Überprüfung und Aufbereitung in die *elexiko*-Stichwortliste überführt.[19] Durch den Vergleich mit Lemmastrecken weiterer Wörterbücher wurde sichergestellt, dass keine wichtigen Stichwörter in *elexiko* fehlen.

1.1.4 Verarbeitung der Sprachressourcen – Programmmodule

Durch den Vergleich der NLP-Tool spezifischen Lemmalisten untereinander und mit der WW-Stichwortliste wird automatisch eine Lemmaliste mit POS-Tags und Frequenzangaben generiert, die mit lexikografischen Ansprüchen korreliert. Die Verarbeitung wurde mit Perl- und Shell-Skripten realisiert, die Ergebnisse der einzelnen Verarbeitungsschritte werden automatisch dokumentiert. Anhand dieser Ergebnisse zeigen sich Spezifika in Bezug auf die Lemmatisierung und das POS-Tagging der einzelnen NLP-Tools, die in Stadler (2013a, 2013b) exemplarisch ausgeführt sind.

Die Erstellung der BLL der nhd. Standardsprache ist auf mehrere aufeinander folgende Schritte verteilt, in denen die Sprachressourcen automatisch verarbeitet werden. Die Module 1 bis 4 verarbeiten die annotierten Korpora und die daraus extrahierten Lemmalisten separat:

1. Extraktion der relevanten Daten (Lemma, POS-Tag, Frequenz) aus den linguistisch annotierten Korpora, Erstellung der NLP-Tool spezifischen „Ur“-Lemmalisten
2. Vereinheitlichung, Bereinigung und Korrektur der Lemmata, Unifikation der Tagsets
3. Abgleich und Alignierung der Lemmalisten mit der WW-Stichwortliste, Erzeugung der NLP-Tool spezifischen Basis- und Restlemmalisten
4. Erzeugung der NLP-Tool spezifischen Eigennamenlisten

Modul 5 übernimmt die Zusammenführung der NLP-Tool spezifischen BLL'en und Eigennamenlisten zu (einer) finalen Version(en).

5. Merging der NLP-Tool spezifischen BLL'en und Eigennamenlisten

19 Die einzelnen Schritte auf dem Weg von der Stichwortkandidatenliste zur *elexiko*-Stichwortliste werden in Schnörch (2005) detailliert beschrieben.

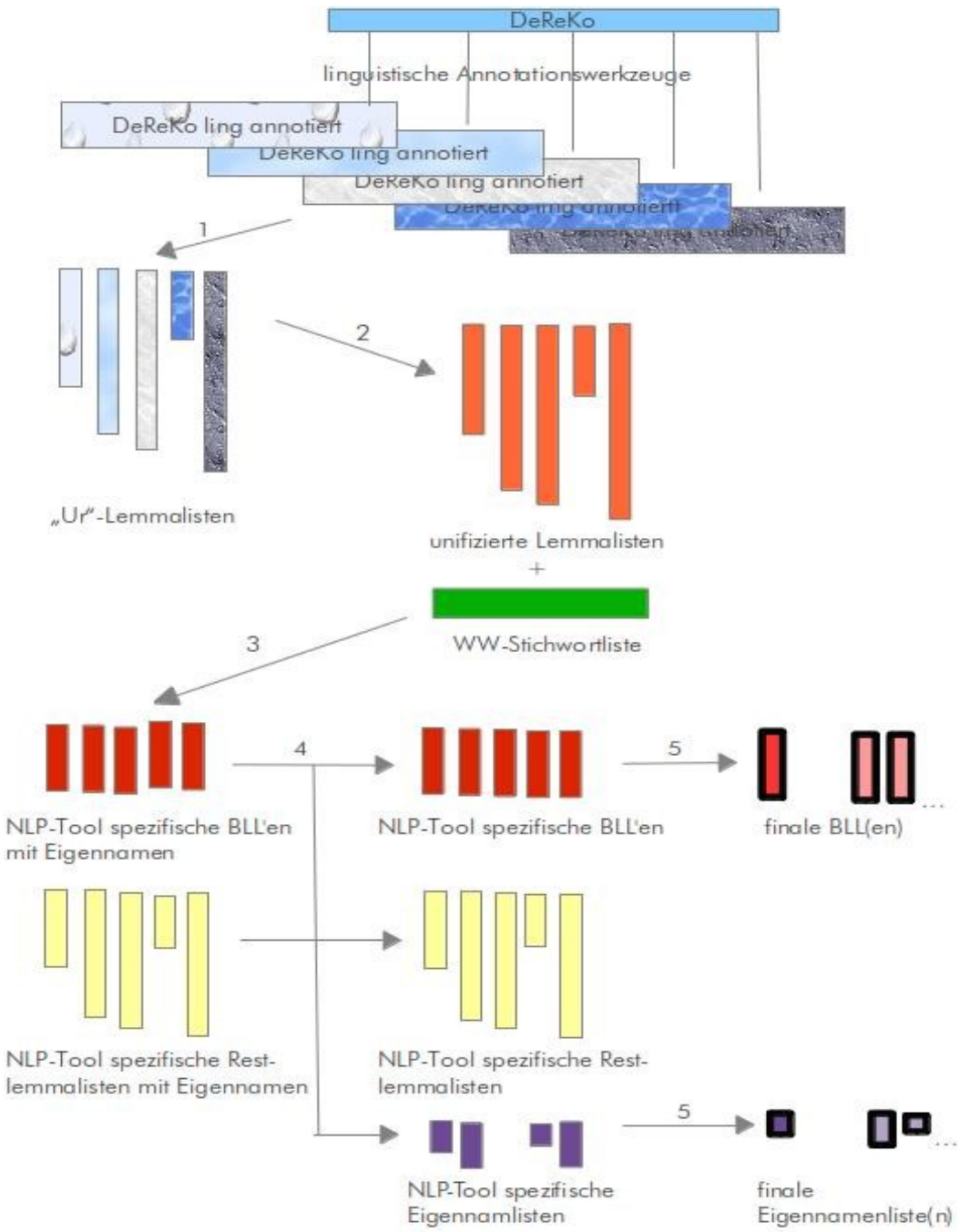

Abb. 5: Die Programmmodule

Die automatisch generierten NLP-Tools spezifischen Lemmalisten sind eher durch die Lemmatisierungskonzepte des NLP-Tools beeinflusst als durch Korpusbesonderheiten. Lemmalisten, die von verschiedenen NLP-Tools mit dem gleichen Korpus erstellt sind, weichen in Bezug auf die Lemmatisierung stärker voneinander ab, als die von einem NLP-Tool mit verschiedenen (gleich großen) Korpora erstellten Listen. Die morphosyntaktische Analyse in der automatischen Sprachverarbeitung impliziert fehlerhafte und zwischen den NLP-Tools divergierende Lemmaformen und POS-Tags, die falsche Zuordnung von Wortformen zu Lemmata, die Nicht-Erkennung von Wortformen, deren Lemmata folglich fehlen oder die sich

aus den Wortformen selbst konstituieren, und die Nicht-Auflösung von Lemma- und POS-Ambiguitäten. Das eigentliche Problem bei der automatischen Erstellung korpusbasierter Lemmalisten stellt nicht die Übernahme von Orthografiefehlern aus dem Korpus dar, sondern das Einfließen von Buchstabenkombinationen, die keine deutschen Lexeme sind, die von den NLP-Tools produzierten Lemmatisierungsfehler, nicht analysierte Wortformen und systematische Abweichungen von der Stichwortkonzeption für Wörterbücher.

Die Programme sind modular aufgebaut, das Korpus und die linguistischen Annotationswerkzeuge können ausgetauscht oder neue Sprachressourcen integriert werden. Der Vergleich mit der WW-Stichwortliste kann auch übersprungen und eine BLL allein durch die Komparation der aus den annotierten Korpora extrahierten NLP-Tool spezifischen Lemmalisten generiert werden. In diesem Fall kann neben der Frequenz nur das Vorkommen einer Lemma-POS-Tag Kombination in mehreren NLP-Tool spezifischen Lemmalisten als Kriterium für seine Richtigkeit und damit für die Aufnahme in die BLL dienen. Bestimmte regelhafte Lemmatisierungsfehler (*Flitterwoche, Kontodatum, Lebensumstand, Landspiel, Spielsachebasar, nächstgut, weltgroß, Minderlohn*) und die Verwendung von flektierten oder satzinitialen Wortformen als zusätzliche Lemmata werden jedoch von sämtlichen hier untersuchten NLP-Tools praktiziert, wodurch Sprachmaterial, das nicht der nhd. Standardsprache angehört, in der BLL stehen würde. Auch führen sämtliche NLP-Tools adjektivisch verwendete Partizipien und Substantivierungen als zusätzliches Lemma neben der verbalen Grundform, unabhängig von Bedeutungsgleichheit oder Lexikalisierung. Im Allgemeinen werden Partizipien in Wörterbüchern unter dem Lemma des Verbs subsumiert, wenn sie keine lexikalisierte Bedeutung haben, die sich von der des Verbs unterscheidet.[20] Da die partizipialen Lemmaformen vieler Verben sehr häufig im Korpus auftreten, würden gerade im vorderen Bereich der BLL zahlreiche unter lexikografischen Gesichtspunkten überflüssige Lemmata stehen *(gegangen, gekauft, verkauft, abbiegend, variierend, zerfließend).*

Die Korpusextraktion der Lemmalisten (Modul 1) verläuft NLP-Tool spezifisch aufgrund der stark abweichenden Granularität der linguistischen Annotationen und der unterschiedlichen XML- bzw. XCES-Auszeichnung der einzeln annotierten Korpora. Die Frequenzen der Lemmata mit dem dazugehörigen POS-Tag werden addiert und in NLP-Tool spezifischen frequenzsortierten Lemmalisten subsumiert. Ein Vergleich der häufigsten Lemmata in den frequenzsortierten NLP-Tool spezifischen „Ur“-Lemmalisten verdeutlicht die Notwendigkeit der Vereinheitlichung von POS-Tags und bestimmten Lemmata (Artikel und Pronomen):

20 Kriterien für die Aufnahme von Partizipien in die Stichwortlisten von Wörterbüchern werden in Erb (2005: 93-94) anhand des Vorgehens bei *elexiko* eingehend erläutert.

glemm			MPT			TreeTagger		
d-	ART	372041251	die	DET	195144842	die	ART	348613119
in	PRA	102571346	der	DET	140798904	,	$,	220936151
und	CON	84379423	in	PREP	100152883	.	$.	209869299
sein	VRB	59777467	.	V	92956679	und	KON	85672890
werden	VRB	41845240	und	CC	85465442	UNKNOWN	NE	78522261
von	PRA	40739636	,	N	77073025	eine	ART	75238859
mit	ADV	35192255	,	V	69769840	in	APPR	66855760
haben	VRB	34950734	.	N	65259304	"	$(	64893662
ein	PRO	33762990	sein	V	64537359	unknown	NN	61548996
an	PRA	31994097	das	DET	63511136	sein	VAFIN	59478241
zu	FUN	31115731	zu	PREP	45016087	@card@	CARD	34578550
für	PRA	26860147	ein	DET	42900667	von	APPR	34547607
er/sie/es/sie	PRO	26382736	werden	V	41814302	mit	APPR	33822577
auf	PRA	25893375	von	PREP	38781798	haben	VAFIN	33506610
nicht	ADV	23019130	die	PRON	36974814	im	APPRART	30947199
ein	ART	21870507	haben	V	35623879	:	$.	29864399
eine	NUM	21323830	,	NE	29851099	werden	VAFIN	29848230
es	PRO	21016363	eine	DET	29293374	für	APPR	27937099
auch	ADV	20274147	mit	PREP	28251627	er|es|sie	PRF	26745646
bei	PRA	19374426	für	PREP	28099826	auf	APPR	24950777
zu	PRA	17978658	sich	PRON	26734882	)	$(	24493758
sie	PRO	17907983	an	PREP	26043773	(	$(	24455221
er	PRO	17644340	auf	PREP	24814732	nicht	PTKNEG	23471015
als	PRA	17589707	es	PRON	23932297	die	PRELS	23210571
sein	PRO	15404538	nicht	ADV	23410488	es	PPER	21552762
nach	PRA	15182829	"	N	21870500	auch	ADV	20529329
…			…			…		

Abb. 6: Frequenzsortierte NLP-Tool spezifische „Ur“-Lemmalisten

Von Modul 2 werden mehrere Schritte zur Vereinheitlichung und Korrektur der Lemmalisten ausgeführt:

- Extraktion von Lemmata, die Sonderzeichen (@,#,www) oder keine alphabetische Zeichen (Zahlen, Satzzeichen) enthalten
- Überführung der Lemmata von *MPT* und *XIP* durch regelbasierte Algorithmen auf eine in der Lexikografie übliche Form[21]
- Vereinheitlichung der Lemmaformen für Pronomen und Artikel mittels einer händisch erstellten Liste
- Unifikation der Tagsets
- Ermittlung des Konfidenzwertes für die Güte der POS-Tags
- Korrektur sehr häufiger falscher Lemmata und POS-Tags mittels einer händisch erstellten Liste
- Generierung von Testlisten: ambige Lemmata (ursprüngliches und unifiziertes Tagset), Tagset mit Häufigkeit der POS-Tags (ursprüngliches und unifiziertes Tagset)

21 Die internen Morphologiekomponenten von *MPT* und *XIP* führen einzelne Morpheme in Komposita auf ihre (vermeintlichen) Grundformen zurück und bilden aus ihnen das Lemma (*haupt frieden hof* 'Hauptfriedhof', *alter grund* 'Altersgründe', *schule abschluss - schulen#abschluss* 'Schulabschluss', *dunkeln#blau* 'dunkelblau', *mit=gliedern#Versammelung* 'Mitgliederversammlung', *fachen#Hochschule* 'Fachhochschule', *an=liegen}heit* 'Angelegenheit'). *MS* stellt die Bestandteile, aus denen Komposita bestehen, in der Lemmaform lediglich durch eine Raute getrennt und den Anfangsbuchstaben als Minuskel dar (haup#fried#hof, dunkel#blau).

glemm			MPT			TreeTagger		
der,die,das	DET	372041251	der,die,das	DET	399454882	der,die,das	DET	350956922
in	PREP	102571346	in	PREP	100152883	und	CONJ	85672890
und	CONJ	84379423	und	CONJ	85465442	ein(e)	DET	75834874
sein	V	59777467	ein(e)	DET	72201801	in	PREP	66870303
werden	V	41846394	der,die,das	PRON	67395083	sein	V	64499919
von	PREP	40739636	sein	V	64537359	werden	V	41819359
mit	PREP	35192255	zu	PREP	45016087	haben	V	35386600
haben	V	34950734	werden	V	41814302	der,die,das	PRON	34973830
ein(e)	PRON	33762990	von	PREP	38781798	von	PREP	34553397
an	PREP	31994097	haben	V	35623879	mit	PREP	33822577
zu	ADV	31115731	mit	PREP	34859900	im	PREP	30947199
für	PREP	26881002	für	PREP	28099826	für	PREP	27949705
sich	PRON	26341487	sich	PRON	26734882	sich	PRON	26745646
auf	PREP	25893375	an	PREP	26043773	auf	PREP	24955025
ein(e)	DET	25776090	auf	PREP	24814732	nicht	ADV	23471015
nicht	ADV	23019130	es	PRON	23932297	es	PRON	21552762
ein(e)	NUM	21329931	nicht	ADV	23410488	auch	ADV	20529329
es	PRON	21016363	auch	ADV	19599208	er	PRON	20437336
auch	ADV	20274147	sie	PRON	19576575	zu	CONJ	19307431
bei	PREP	19374426	er	PRON	18331833	als	CONJ	17354281
sein	PRON	18596588	als	CONJ	17870326	am	PREP	16497099
zu	PREP	17978658	bei	PREP	17368189	sein	PRON	15761347
sie	PRON	17907983	sein	PRON	15559970	dies(e,er,es)	PRON	15329952
er	PRON	17644340	dies(e,er,es)	PRON	15421438	bei	PREP	14377464
als	CONJ	17589707	dass	CONJ	14242391	nach	PREP	14136194
nach	PREP	15182829	ihr	PRON	14118537	können	V	14045367
dies(e,er,es)	PRON	14816130	können	V	14103454	sie	PRON	13647883
aus	PREP	14772649	Jahr	NN	11794241	ihr	PRON	13577927
…			…			…		

Abb. 7: Frequenzsortierte NLP-Tools spezifische Lemmalisten nach Durchlaufen des Moduls 2

Modul 3 übernimmt den Abgleich der NLP-Tool spezifischen Lemmalisten mit der WW-Stichwortliste. Die BLL soll keine Lemmata enthalten, die es im Deutschen nicht gibt. Lemmatisierungsfehler, Orthografiefehler oder fremdsprachliche Belegstellen führen ebenso zu Lemmata in den Lemmalisten, die in keinem Wörterbuch der nhd. Standardsprache stehen, wie die vom lexikografischen Standard abweichenden Lemmatisierungskonzepte der NLP-Tools. Die in der WW-Stichwortliste nicht enthaltenen Partizipien und Substantivierungen werden automatisch mit den in der WW-Stichwortliste verzeichneten Grundformen aligniert und die Frequenzen zu denen der Grundform addiert. Übersteigt die Frequenz von Substantivierungen (*Aufständische, Verantwortliche, Hinterbliebene, Nachsehen*) oder adjektivisch gebrauchten Partizipien (*abschließend, nachfolgend, aufregend, rauschend, eingeschlossen*) die Frequenz der Grundform, werden die abgeleiteten Formen zusätzlich in die BLL aufgenommen. Der sprachliche Usus soll sich in der BLL widerspiegeln. Von Modul 3 werden folgende Schritte beim Abgleich der NLP-Tool spezifischen Lemmalisten mit der WW-Stichwortliste ausgeführt:

- exakter Stringvergleich
- Mapping der orthografischen Varianten auf die Standardschreibweise
- Alignierung ss-ß
- Alignierung Paritzip1
- Alignierung Partizip2
- Alignierung Substantivierungen
- Alignierung Pluraliatantum

Durch den Abgleich mit der WW-Stichwortliste entstehen NLP-Tool spezifische BLL'en, die lexikografisch validierte Lemmata enthalten. Daneben werden NLP-Tool spezifische Restlemmalisten gebildet, in die diejenigen Lemmata einfließen, die nicht in der WW-Stichwortliste verzeichnet sind und die nicht aligniert werden können. Die BLL wächst in einem iterativen Prozess: durch die händische Bearbeitung der frequenzsortierten Restlemmalisten werden neue Lemmata verifiziert und in die WW-Stichwortliste integriert oder neue Alignierungsalgorithmen implementiert. Die betroffenen Lemmata stehen nach dem nächsten Programmlauf nicht mehr in der Restlemmaliste, sondern in der BLL.

Abb. 8: Programmmodul 3

In der Restlemmaliste befinden sich sowohl Lemmata, die nach einer händischen Überprüfung oder der Implementierung weiterer Alignierungsalgorithmen in die BLL einfließen, als auch sprachliches Material, das unter lexikografischen Gesichtspunkten nicht in die BLL übernommen wird. Hauptsächlich im vorderen Bereich der frequenzsortierten Restlemmalisten stehen viele Lemmata, die spezifische Lemmaformen der einzelnen NLP-Tools darstellen (*Fall|Falle, BERLIN, Usa, sondern auch*), falsche Lemmatisierungen *(hamburger, Salzbürger, unbedingen, unverändern, Deshalb, montag, gottloben*), Orthografiefehler (*wiedersprechen, Prophezeihung, schießlich*) und Wortformen, die über die Alignierung noch nicht auf ihre Grundform zurückgeführt wurden (*Große, Ganze, Gleichzeitig, Seinerzeit*).

Die verbleibenden lexikografisch interessanten Lemmata in den Restlemmalisten verteilen sich überwiegend auf Komposita, Neologismen und Fremdwörter (*Abfallsünder, Airshow, alternativlos, Aquafitness, Ärztemangel, Befristungsgesetz, Bergleute, berufsübergreifend, Bitfolge, Boyband, Brustwirbelsäule, deinstallieren, Ferienplanung, googeln, Handelsstreitigkeit,*

Hedgefonds, Hochseilgarten, Internetblase, jammen, Kleingewerbetreibende, Königsposition, Kopftuchträgerin, Kostenfalle, Mentorenprogramm, Mixtape, Onlinebroker, Opferdiskurs, Orchesterhaus, Priesterstand, Radiomultikulti, Ratingsystem, Ravekultur, Rolltor, Romafamilie, Schiedsrichterkosten, Seniorenrunde, spacig, Stabmixer, Trefferwahrscheinlichkeit, Trinkwasserprojekt, Voting, Weblink, Wildschweinbestand, Wohnungslose), daneben sind Herkunftsbezeichnungen (*Lampertheimer, Ludwigshafener, Yorker, tschetschenisch*), Eigennamen (*Carlos, Black, Romanshorn, Strobl, Hillary*) und Movierungsvarianten (*Neugeborene, Geschworene, Kriminalbeamte*) vertreten. Durch die sukzessive händische Validierung der lexikografisch relevanten häufigen Lemmata aus den Restlemmalisten wird die BLL aktualisiert und an die Lexik der sich aus der Korpuszusammensetzung ergebenden Themen angepasst.

Modul 4 extrahiert die Eigennamen aus den NLP-Tool spezifischen BLL'en und Restlemmalisten und fasst sie in NLP-Tool spezifischen Eigennamenlisten zusammen. Einen zusätzlichen POS-Tag für Eigennamen vergeben *glemm, MPT, TreeTagger* und *XIP.* Die Anzahl der Eigennamen und die Eigennamen selbst sind zwischen den NLP-Tools stark different. Für die Aufnahme in die Eigennamenliste 1.0 muss ein Eigenname von mindestens zwei NLP-Tools als solcher erkannt werden, eines der beiden NLP-Tools, die den Eigennamen verzeichnen, muss *glemm* (28.817 Eigennamen) oder *TreeTagger* (36.769 Eigennamen) sein. Dadurch wird die Anzahl von 4.402.071 Eigennamen bei *MPT* bzw. 6.277.930 Eigennamen bei *XIP* erheblich reduziert. *MPT* und *XIP* annotieren nicht analysierbare nominale Tokens (*Reissnägel, Baltzerdie, Ballstopper, Assisstent*) und längere Komposita (*Affekt-Kontroll-Training, Ballettvorbereitung*) sehr häufig als Eigennamen, obwohl es sich um Tokenisierungs-, Orthografiefehler oder Gattungsnamen handelt. Über zusätzliche manuell erstellte Listen, die nur Nomen enthalten, die ausschließlich als Eigennamen fungieren, werden nicht erkannte (und folglich mit dem POS-TAG NN ausgezeichneten) Eigennamen aus den Lemmalisten extrahiert und der POS-Tag bei der Integration in die Eigennamenliste auf NE umgesetzt. Mit Mustererkennung (*straße, gasse, bach, hausen*) werden weitere Eigennamen in den Lemmalisten gefunden, die den POS-Tag NN führen. Die Mustererkennung wird nur auf die Restlemmaliste angewandt, in der BLL sind überwiegend Gattungsnamen mit den Mustern enthalten (*Umgehungsstraße, Schnellstraße, Wasserstraße*). Die nach diesen Kriterien generierte Eigennamenliste 1.0 verzeichnet 68.892 Einträge, eine Eigennamenliste, die ohne Mustererkennung erstellt wird, verzeichnet 39.261 Einträge.

Der Abgleich der NLP-Tool spezifischen BLL'en oder Eigennamenlisten in Modul 5 garantiert ein Maximum an Informationen zu einem Lemma, dessen POS-Tag(s) und Häufigkeit im Korpus. Die NLP-Tool spezifischen Lemmalisten unterscheiden sich neben den Lemmata auch in den POS-Tags zu den einzelnen Lemmata und den akkumulierten Frequenzen. Es gibt keine NLP-Tool spezifische Lemmaliste, die für Lemmata und POS-Tags hundertprozentig verlässliche Ergebnisse (Precision) und eine komplette Abdeckung der Wortformen im Korpus (Recall) bietet. Eine BLL der nhd. Standardsprache variiert nicht nur mit der Korpusgrundlage, sondern vor allem mit den ausgewählten NLP-Tools, die bei der linguistischen Annotation des Korpus Verwendung finden, sowie dem Aufwand der lexikografischen Nachbearbeitung.

Das Aufnahmekriterium für ein Lemma in die BLL ist sein Vorhandensein in der WW-Stichwortliste. Ausschlaggebend für die Übernahme eines POS-Tags für ein bestimmtes Lemma in die BLL ist das Vorkommen des POS-Tags mit dem betreffenden Lemma in mehreren NLP-Tool spezifischen BLL'en. Weichen die POS-Tags für ein Lemma in den NLP-Tool spezifischen Lemmalisten voneinander ab, werden die POS-Tags des NLP-Tools mit einem hohen Konfidenzwert für das POS-Tagging übernommen. Der Konfidenzwert für das POS-Tagging der einzelnen NLP-Tools wird anhand des manuell ausgezeichneten TIGER-Korpus ermittelt.[22] In die Berechnung des Konfidenzwertes fließt ein, wie viele der POS-Tags der NLP-Tool spezifischen Lemmalisten jeweils mit den POS-Tags für die Lemmata aus dem TIGER-Korpus identisch sind und wie viele POS-Tags davon abweichen.

Während des Vergleichs der NLP-Tool spezifischen BLL'en wird ein Mittelwert aus den teilweise stark abweichenden absoluten Frequenzen für ein Lemma mit dem zugehörigen POS-Tag berechnet, der in der finalen BLL verzeichnet ist. Als finale BLL wird diejenige Version gewählt, die aus dem Vergleich der vier BLL'en von *Glemm, MPT, MS* und *TreeTagger* entsteht (die Annotationen von XIP waren zum Entstehungszeitpunkt der BLL 1.0 nicht verfügbar). Auch die Zusammenführung von nur zwei oder drei NLP-Tool spezifischen BLL'en oder Eigennamenlisten ist in Modul 5 möglich und wird zu Vergleichszwecken durchgeführt.

Wünschenswert wäre eine automatische lexikalische Akquisition, die lexikografischen Standards entspricht. Um dieses Ziel zu erreichen, ist zum einen die Einbindung des lexikografischen Wissens, das in Wörterbüchern steckt, in den automatischen Verarbeitungsprozess der NLP-Tools erforderlich. Lexikografische Informationen über Abweichungen von generellen Lemmatisierungsregeln (z.B. Pluraliatantum, lexikalisierte Partizipien und Superlative) können mit Listen in die Analyse eingebunden werden - die lexikografischen

22 TIGER-Korpus: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

Informationen, die in den Listen kodiert sind, können durchaus auch a priori korpusbasiert gewonnen werden. Zum anderen sollten die Informationen, die die NLP-Tools aus einer Analyse des Satzkontextes gewinnen, in die Auflösung von morphosyntaktischen Ambiguitäten zurückfließen und nicht nur der Erkennung syntaktischer Funktionen dienen. Nicht immer ist die Disambiguierung heterolemmatischer oder mehreren Wortarten zugehöriger Wortformen über die Einbeziehung des Kontextes möglich, doch wäre das in dem spezifischen Zusammenhang relevante Lemma oder der relevante POS-Tag in vielen Fällen zu bestimmen. Für eine Verbesserung der Ergebnisse der NLP-Tools in Bezug auf die Lemmatisierung und das POS-Tagging sind die Optimierung der Lexika und morphologischen Analysen sowie die Auswertung des Kontextes relevant. Die Möglichkeiten und Grenzen der automatischen Erstellung von korpusbasierten, lexikografisch validen Lemmalisten korrelieren mit dem Aufwand, der in die Implementierung der NLP-Tools investiert wird. Die BLL kombiniert die korpusbasiert ermittelten linguistischen Daten mit lexikografischem Expertenwissen und bietet eine unter lexikografischen Gesichtspunkten optimierte Abbildung der schriftlichen nhd. Standardsprache.

1.2 Varietätenlemmalisten

Im Fokus der achtmonatigen Projektverlängerung stand die Erstellung und der Vergleich von korpusbasierten Varietätenlemmalisten. Die Stratifikation von DeReKo legt eine Untersuchung der regionalen Merkmale der schriftlichen Standardsprache nahe (vgl. 1.1.3). Das Deutsche ist eine plurizentrische Sprache, in den drei Ländern Deutschland, Österreich und der Schweiz gelten mitunter unterschiedliche national definierte Sprachstandards in Bezug auf die Orthografie, Lexik und Aussprache.[23] Das Deutsche ist zudem eine pluriareale Sprache, denn auch innerhalb der deutschsprachigen Länder sind mehrere Sprachlandschaften zu differenzieren. Berend (2005) beschreibt regionale Gebrauchsstandards, die sich vom Dialekt durch ihre Akzeptanz in formellen Gesprächssituationen unterscheiden.

Die räumliche Aufteilung eines Sprachgebiets wird üblicherweise in Sprachatlanten dargestellt.[24] Festgelegt werden die sprachlichen Grenzen durch unterschiedliche linguistische Kriterien (Aussprache, grammatische Konstruktionen, Wortschatz) und auf unterschiedlichen soziolinguistischen Ebenen (Dialekte, Umgangssprache, Regionalsprache, Mundarten, Aussprache des Schriftdeutschen). Gemeinsam ist den Sprachatlanten die gesprochene Sprache als Untersuchungsgegenstand. In Wörterbüchern werden die Besonderheiten in der Lexik, die

23 Ammon (1995), Ammon/Bickel/Ebner et al. (2004), Clyne (1992).
24 Z.B. Eichhoff (1977-2000), Elspaß/Möller, Mitzka/Schmitt (1951-1980).

die Schriftsprache von Österreich und der Schweiz auszeichnen, erfasst. Neben der Kennzeichnung einzelner Stichwörter in allgemeinsprachlichen Wörterbüchern als in ihrer Verwendung regional beschränkt[25], gibt es Monographien die den spezifischen Wortschatz der Österreich und der Schweiz verzeichnen.[26] Bundesdeutsche Besonderheiten hingegen, die für das gesamte Sprachgebiet der BRD charakteristisch sind, jedoch nicht in der Schweiz oder Österreich gebräuchlich sind, werden in Wörterbüchern üblicherweise nicht extra markiert. Der rein bundesdeutsche Wortschatz wird in keinem Wörterbuch thematisiert.

Für die zahlreichen in DeReKo vertretenen Zeitungen aus Deutschland, Österreich und der Schweiz wurde ermittelt, inwieweit die spezifische Lexik der Zeitungen mit ihrer räumlichen Herkunft in Zusammenhang steht. Mit Algorithmen aus der Dialektometrie wurde zunächst untersucht, ob sich im Zeitungskorpus nationale Grenzen oder Dialektgebiete im standardsprachlichen Wortschatz widerspiegeln. Der bayrische oder alemannische Dialektraum beispielsweise schließt Gebiete aus Deutschland und Österreich bzw. der Schweiz mit ein und unterscheidet sich von Sprachgebieten, die in der gleichen Nation liegen.

Die Dialektometrie ist ein „numerisch-mathematisches Verfahren zur Auswertung und Komprimierung von in Sprachatlanten enthaltenen geolinguistischen Daten“ (Goebl 2002: 179). Ihre Verfahren werden meist auf phonetische Daten angewandt, zwischen denen ein Grad der Verschiedenheit berechnet werden kann. Die Zuordnung äquivalenter Lemmata (*Tomate – Paradeiser*) ist nicht möglich. Um die binäre Variation (identisch / nicht identisch) von lexikalischen Daten zu aggregieren ist der Relative Identitätswert geeignet (Zastrow 2011: 43-45). Der Relative Identitätswert (RIW) stellt die Relation zwischen gleichen (Koidentität, KOI) und ungleichen Entitäten (Kodifferenz, KOD) zweier Messreihen dar:

$$RIW_{j,k} = 100 \cdot \frac{\sum_1^i KOI_{j,k}^i}{\sum_1^i KOI_{j,k}^i + \sum_1^i KOD_{j,k}^i} \qquad (1)$$

Als Messreihen werden die Lemmalisten der einzelnen Zeitungen angesetzt.[27] Als Koidentität gilt das Vorhandensein eines Lemmas in den beiden untersuchten Lemmalisten, als Kodifferenz das Fehlen eines Lemmas in der anderen Lemmaliste. Aus den berechneten Relativen Identitätswerten zwischen den einzelnen Zeitungen wird automatisch eine Matrix generiert, die die Ergebnisse verzeichnet. In einer Reihe von Versuchen hat sich gezeigt, dass nur Zeitungen mit übereinstimmenden Jahrgängen verglichen werden können, um zufriedenstellende

25 Z.B. Retti (1999), Bickel/Hofer (2002).
26 Z.B. Bickel/Landolt (2012), Ebner (2009), Meyer (1989).
27 Die Varietätenlemmalisten werden mit den linguistischen Annotationen von DeReKo durch *Machinese Syntax* erstellt.

Resultate zu erzielen. Der Bezug zum Zeitgeschehen in Zeitungstexten schlägt sich in starkem Maße in der Lexik nieder und überlagert weitere Faktoren wie die Herkunft der Zeitung, ihre regionale oder überregionale Ausrichtung, die Gewichtung der Ressorts, Themen und Rubriken oder ihre Seriosität.

In der Matrix in Tabelle 1 werden die Ergebnisse für sechs Zeitungen aus der BRD und Österreich und zwei Zeitungen aus der Schweiz dargestellt (für die Schweiz gibt es für einzelne Jahrgänge nicht mehr als zwei Zeitungskorpora in ausreichender Größe). Für die BRD und für Österreich werden die fünf höchsten Werte, für die Schweiz der höchste Wert pro Zeile, die sich aus der Berechnung des Relativen Identitätswerts mit anderen Zeitungen ergeben, fett markiert. Zur Berechnung des Relativen Identitätswerts wird in den unterschiedlich langen Lemmalisten der einzelnen Zeitungen jeweils nur für die ersten 50.000 Lemmata überprüft, ob sie in der Lemmaliste der anderen Zeitung enthalten sind. Die unterschiedlichen Größen der einzelnen Zeitungskorpora werden dadurch ausgeglichen sowie die zahlreichen niedrigfrequenten Token, die keine deutschen Wörter bezeichnen, bei der Berechnung ignoriert. Die höchsten Werte für die relative Identität sind fast immer innerhalb der gleichen nationalen Varietät zu finden:

	BM	FAZ	FR	MM	NN	TAZ	DP	KL	NK	SN	TT	VA	ST	ZT	
Berliner Morgenpost	**100**	**55**	**60**	**52**	**55**	**55**	46	44	39	48	44	45	41	45	
Frankfurter Allgemeine	**55**	**100**	**65**	49	**57**	**57**	**51**	42	37	49	43	44	41	50	
Frankfurter Rundschau	**60**	**65**	**100**	**55**	**60**	**61**	50	46	39	50	46	47	44	50	D
Mannheimer Morgen	**52**	**49**	**55**	**100**	**50**	**45**	41	43	39	44	44	44	42	41	
Nürnberger Nachrichten	**55**	**57**	**60**	**50**	**100**	**56**	49	45	39	49	45	45	41	47	
taz	**55**	**57**	**61**	45	**56**	**100**	**50**	42	37	47	42	42	39	49	
Die Presse	46	**51**	**50**	41	49	**50**	**100**	**51**	46	**60**	**51**	**50**	39	48	
Kleine Zeitung	44	42	46	43	45	42	**51**	**100**	**53**	**55**	**55**	**52**	41	42	
Neue Kronen Zeitung	39	37	39	39	39	37	**46**	**53**	**100**	**51**	**49**	**47**	36	37	Ö
Salzburger Nachrichten	48	49	50	44	49	47	**60**	**55**	**51**	**100**	**55**	**54**	41	47	
Tiroler Tageszeitung	44	43	46	44	45	42	**51**	**55**	**49**	**55**	**100**	**55**	41	42	
Vorarlberger Zeitung	45	44	**47**	44	45	42	**50**	**52**	**47**	**54**	**55**	**100**	43	43	
St. Galler Tagblatt	41	41	44	42	41	39	39	41	36	41	41	43	**100**	**52**	CH
Zürcher Tagesanzeiger	45	50	50	41	47	49	48	42	37	47	42	43	**52**	**100**	
			Deutschland						Österreich				Schweiz		

Tabelle 1: Verteilung des Relativen Identitätswertes auf Zeitungen aus Deutschland, Österreich und der Schweiz

In einem weiteren Versuch wurde anhand von elf Zeitungen aus Deutschland untersucht, ob sich Dialektgrenzen innerhalb eines Landes in der Lexik der schriftlichen Standardsprache

abzeichnen. Es stellte sich jedoch heraus, dass die regionale Differenzierung des in den Zeitungen befindlichen Wortschatzes innerhalb eines Landes nicht ausreicht, um regionale Gruppierungen zu erreichen. Zwar haben z.B. die *Nürnberger Nachrichten* und die *Nürnberger Zeitung* einen hohen relativen Identitätswert, doch weist *die taz* einen höheren Identitätswert mit der *Zeit* oder anderen Zeitungen auf als mit der *Berliner Zeitung.*

Zur Überprüfung der These, dass sich lexikalische Besonderheiten in der schriftlichen Standardsprache in stärkerem Maße nur zwischen den Nationen und nicht innerhalb einer Nation zeigen, wurde ein Algorithmus aus der Statistik, ein Clusterverfahren, angewandt. Das Ziel von Clusterverfahren ist die Zusammenfassung von Objekten zu Gruppen (Cluster), wobei es eine möglichst große Homogenität innerhalb eines Clusters und eine möglichst große Heterogenität zwischen den Clustern zu erreichen gilt. Das Clusterverfahren k-Means gehört zu den partitionierenden, globalen Clusterverfahren mit exakter Zuordnung, das die gewählte Anfangspartition der zu clusternden Objekte durch Neuzuordnung schrittweise verbessert.[28] Abstandsmaß in der vorliegenden Implementierung von k-Means ist die Euklidische Distanz:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2)$$

In die Berechnung der Cluster fließt die Häufigkeit der einzelnen Lemmata ein. Die Matrix, die als Berechnungsgrundlage dient, beinhaltet die Frequenzen der Lemmata in den Lemmalisten der einzelnen Zeitungen. Die besten Ergebnisse werden erzielt, wenn in den Lemmalisten nur nominale Lemmata oder alternativ auch adjektivische und verbale Lemmata stehen:

BM	**DP**	**FAZ**	FR	KL	MM	NK	NN	RZ	SN	ST	TAZ	TT	VA	ZEIT	**ZT**	
0	**20**	**18**	10	0	20	20	0	20	13	6	20	15	20	20	**19**	Aaa#N
0	**5**	**0**	0	0	5	5	0	5	0	0	5	0	5	5	**0**	Aaaah#N
0	**8**	**0**	0	0	8	8	0	8	0	0	8	0	8	8	**0**	Aaah#N
0	**5**	**0**	5	0	5	5	0	5	6	0	5	0	5	5	**5**	Aaas#N
13	**15**	**0**	0	27	15	15	0	15	0	0	15	61	15	15	**0**	Aab#N
0	**0**	**0**	0	0	0	0	0	0	0	0	0	0	0	0	**15**	Aabach#N
0	**0**	**0**	9	0	0	0	0	0	0	0	0	0	0	0	**0**	Aac#N
0	**7**	**6**	230	0	7	7	0	7	0	91	7	0	7	7	**0**	Aach#N
0	**0**	**0**	0	0	0	0	0	0	0	15	0	0	0	0	**0**	Aachbrücke#N
0	**0**	**0**	6	0	0	0	0	0	0	0	0	0	0	0	**0**	Aache#N
0	**0**	**0**	5	0	0	0	0	0	0	0	0	0	0	0	**0**	Aacheblick#N
258	**85**	**388**	421	84	85	85	86	85	100	80	85	84	85	85	**63**	Aachen#N
48	**10**	**187**	148	0	10	10	46	10	11	10	10	0	10	10	**0**	Aachener#N
0	**0**	**0**	5	0	0	0	0	0	0	0	0	0	0	0	**0**	Aachenerin#N
0	**0**	**0**	0	0	0	0	0	0	0	8	0	0	0	0	**0**	Aachstrasse#N
0	**0**	**0**	0	0	0	0	0	0	0	12	0	0	0	0	**0**	Aachsäge#N
…																…

Tabelle 2: Matrix mit den Frequenzen der nominalen Lemmata der einzelnen Zeitungen zur Berechnung der Euklidischen Distanz. Die gewählten Ausgangsclusterzentren sind fett markiert.

28 Eine formale Darstellung des Algorithmus zeigen Manning/Schütze (1999: 516).

Dem in Perl implementierten Algorithmus können als Parameter sowohl die Anzahl der zu bestimmenden Cluster als auch die Ausgangsclusterzentren übergeben werden. Mittels der Berechnung der Euklidischen Distanz der Vektoren der verbleibenden Zeitungen zu den Vektoren der Ausgangsclusterzentren werden diese demjenigen Clusterzentrum zugeordnet, zu dem die Euklidische Distanz am geringsten ausfällt. Wurden die Zeitungen auf die verschiedenen Clusterzentren verteilt, erfolgt im nächsten Schritt eine Neuberechnung des Vektors der Clusterzentren, der sich aus dem Mittelwert der im Cluster befindlichen Vektoren der Zeitungen ergibt. Das Clusterzentrum eines Clusters (der Zentroid) ist nun nicht mehr identisch mit einer der Zeitungen (außer im Falle eines alleinstehenden Zentroids). In der folgenden Iteration werden die Zeitungen erneut dem Clusterzentrum zugeordnet, zu dem die Euklidische Distanz am geringsten ausfällt. Der Algorithmus wird beendet, wenn die Cluster stabil sind, d.h., dass keine Vertauschungen mehr vorgenommen werden, und sich jede Zeitung in dem Cluster befindet, mit dem sie die maximale Ähnlichkeit in den lexikalischen Daten aufweist. Im Verlauf der Berechnung der Cluster kann es durchaus passieren, dass ein Ausgangsclusterzentrum einem anderen Cluster zugewiesen wird. Mit den drei in Tabelle 2 fett markierten Ausgangsclusterzentren (*Die Presse, FAZ, Zürcher Tagesanzeiger*) verteilen sich die für die Berechnung ausgewählten Zeitungen folgendermaßen auf die berechneten Cluster:

Die Presse:	Die Presse, Kleine Zeitung, Neue Kronen Zeitung, Salzburger Nachrichten, Tiroler Tageszeitung, Vorarlberger Zeitung
FAZ:	Berliner Morgenpost, FAZ, Frankfurter Rundschau, Mannheimer Morgen, Nürnberger Nachrichten, Rhein Zeitung, taz, die Zeit
Zürcher Tagesanzeiger:	St. Galler Tagblatt, Zürcher Tagesanzeiger

Stammen die Anfangsclusterzentren aus jeweils einer anderen Nation, werden die Zeitungen aufgrund ihrer Lexik durch k-Means wie erwartet in die richtigen Cluster eingereiht. Auch mit der Wahl von vier Clusterzentren wird eine mit den Erwartungen konforme Einteilung erreicht:

Salzburger Nachrichten:	Die Presse, Kleine Zeitung, Neue Kronen Zeitung, Salzburger Nachrichten, Tiroler Tageszeitung, Vorarlberger Zeitung
FAZ:	Berliner Morgenpost, FAZ, Frankfurter Rundschau, Nürnberger Nachrichten, taz, die Zeit
Nürnberger Nachrichten:	Mannheimer Morgen, Rhein Zeitung
+Zürcher Tagesanzeiger:	St. Galler Tagblatt, Zürcher Tagesanzeiger

K-Means erweist sich im internationalen Vergleich der Zeitungen selbst mit der Wahl von Ausgangsclusterzentren, die alle aus einer Nation stammen, als ein von der vorgegebenen Einteilung relativ unabhängiger Algorithmus. Die nationalen Grenzen treten auch hier in den berechneten Clustern hervor, doch kommen neben der regionalen Komponente in der Lexik die

Gewichtung der Ressorts, Themen und Rubriken, die Seriosität und das Verbreitungsgebiet der Zeitung stärker zum Tragen:

Neue Kronen Zeitung:	Die Presse, Kleine Zeitung, Neue Kronen Zeitung, Salzburger Nachrichten, Tiroler Tageszeitung
Salzburger Nachrichten:	FAZ, Frankfurter Rundschau, Nürnberger Nachrichten, taz, die Zeit, Zürcher Tagesanzeiger
Vorarlberger Zeitung:	Berliner Morgenpost, Mannheimer Morgen, Rhein Zeitung, St. Galler Tagblatt, Vorarlberger Zeitung

Die Anwendung von k-Means auf Zeitungen, die alle aus Deutschland stammen, zeigt hingegen wie die Berechnung des Relativen Identitätswerts kein klares Bild. Die Teilnehmer an einem Cluster erweisen sich mit wechselnden Anfangsclusterzentren als in hohem Maße inkonsistent, nur selten befinden sich zwei Zeitungen bei mehreren Durchläufen mit veränderten Anfangsclusterzentren im selben Cluster wieder. Innerhalb einer Nation spielt der regionale Einfluss auf die Lexik in der schriftlichen Standardsprache von Zeitungen im Vergleich zu weiteren Faktoren keine messbare Rolle.

In einer weiteren Untersuchung wurden die Zeitungen nach Ländern gebündelt und frequenzsortierte Lemmalisten generiert, die den Standardwortschatz (der Zeitungssprache) einzelner Nationen beinhalten. Durch den Vergleich der nationalen Lemmalisten untereinander werden Varietätenlemmalisten erzeugt, die entweder Lemmata enthalten, die in den Lemmalisten einer anderen Standardvarietät gar nicht verzeichnet sind, oder Lemmata, die sehr viel häufiger auftreten als dieselben Lemmata in der Lemmaliste des anderen Landes.

Ö unikal		Ö Quotient Ö/D > 5		D unikal		D Quotient D/Ö > 5	
Sn#N	22009	Vn#N	1963	Lampertheim#N	11273	Köpenick#N	602
dornbirn#V	6650	Hohenems#N	1013	Bürstadt#N	10127	Vfr#N	577
Spittal#N	6399	Lustenau#N	989	Viernheim#N	8889	**Oberbürgermeisterin#N**	500
Bludenz#N	6245	**Jänner#N**	631	Hanau#N	8572	Biblis#N	487
Gak#N	6089	Weiz#N	576	Hofheim#N	6426	Walldorf#N	344
Lask#N	5064	Dornbirn#N	509	Offenbacher#N	5729	Fag#N	334
Wolfsberg#N	4993	**Landesrat#N**	507	**Sonnabend#N**	5578	Pankow#N	279
Jfrei#N	4564	Steirer#N	483	Ladenburg#N	5184	Kasseler#N	275
Hartberg#N	4057	Schwarzach#N	443	**Ortsbeirat#N**	4358	Wedding#N	273
Villacher#N	4037	Feldkirch#N	392	Hanauer#N	4265	Potsdamer#N	259
Kapfenberg#N	3921	Leoben#N	386	Schriesheim#N	3817	Wentz#N	251
Lkh#N	3452	**Pfarre#N**	360	**Stadtverordnete#N**	3608	Wiesbadener#N	235
Kac#N	3436	Sausgruber#N	357	Ndash#N	3603	Fv#N	230
Voitsberg#N	3124	Lienz#N	338	Lampertheimer#N	3558	Darmstädter#N	229
rankweilen#V	2934	Velden#N	336	Viernheimer#N	3453	**Landrat#N**	223
Wolfurt#N	2804	Halleinen#N	326	lhe#N	3358	**Bundestagsabgeordnete#N**	221
heurig#A	2760	Osttirol#N	304	Oberursel#N	3350	**Bündnisgrüne#N**	201

...

Tabelle 3: Lemmalistenvergleich Österreich - Deutschland 1999
D = Berliner Morgenpost, FAZ, Frankfurter Rundschau, Mannheimer Morgen
Ö = Kleine Zeitung, Neue Kronen Zeitung, Salzburger Nachrichten, Vorarlberger Nachrichten
Lemmata, die keine Kürzel, Eigennamen oder Herkunftsbezeichnungen bezeichnen, sind fett markiert

In den aus den Korpora gewonnenen Daten sind zahlreiche Kürzel und Eigennamen enthalten, die von *Machinese Syntax* als normale Nomen getaggt werden. *Machinese Syntax* verfügt über keine Eigennamenerkennung. Der Verwendung eines NLP-Tools, welches Eigennamen erkennt, und das Aussortieren der Eigennamen ist bei der Erstellung von Varietätenlemmalisten jedoch nicht angeraten, da die NLP-Tools unbekannte Wörter, und damit häufig auch regionale Besonderheiten aus Österreich und der Schweiz (*Jänner, Pfarre*), als Eigennamen verzeichnen. Eine händische Durchsicht der erstellten Varietätenlemmalisten ist zur Erkennung des regional markierten Wortschatzes unerlässlich.

Der für eine Nation korpusbasiert ermittelte charakteristische Standardvarietätenwortschatz wurde mit Einträgen in Wörterbüchern der österreichischen und schweizerischen nationalen Besonderheiten verglichen, die beide in der Duden Reihe erschienen sind. Die völlig überarbeitete vierte Auflage des Wörterbuchs von Ebner (2009) *Wörterbuch des österreichischen Deutsch (Wie sagt man in Österreich?)* erfasst die Standardsprache, Umgangssprache und mundartnahe Sprache. Das *Schweizerhochdeutsch* von Bickel/Landolt (2012) ist ein Wörterbuch der Standardsprache in der deutschen Schweiz. Durch den Vergleich des deutschen, österreichischen und schweizerischen Wortschatzes der korpusbasierten Lemmalisten wird deutlich, dass auch zahlreiche bundesdeutsche Standardvarietäten im Korpus auftreten, für die es jedoch kein selbständiges Wörterbuch gibt.

<u>Vergleich Österreich – Deutschland</u>

Österreich Duden:

Konsumentenschutz, Pensionistenverband, emotionell, fix, Wissenschafter, akkordieren, niederstoßen, erfangen, Lokalaugenschein, …

Österreich nicht im Duden:

Nichtigkeitsbeschwerde, Krankenscheingebühr, parteifrei, terminisieren, Pannenstreifen, Nachbesetzung, Ministeranklage, zinsenlos, zinsenfrei, demokratie-politisch, verhäuten, Sektionsobmann, Wirtschaftspolizei, …

Nur in Deutschland gebräuchlich:

Bagatelldelikt, Metropolregion, Seelsorgeeinheit, proppenvoll, Polizeiverordnung, zinsfrei, zinslos, Rentenversicherungsträger, Einwohnermeldeamt, bürgerschaftlich, terminieren, kreisfrei, prozentual, …

Vergleich Schweiz – Deutschland

Schweiz Duden:

Kantonsrat, Grillplausch, Gemeindeammann, Festwirtschaft, Initiant, Chefredaktor, Grümpelturnier, Medienorientierung, parkieren, konkurrenzieren, grillieren, polysportiv, lüpfig, urchig, eindrücklich, rural, brandmager, abwechslungsweise, …

Schweiz nicht im Duden:

Katastrophenhund, Zugskomposition, Durchgangszentrum, hospitalisieren, statutarisch, Alterszentrum, Altersturnen, arbeitsmarktlich, Kindergärtler, Chrabbelfiir, Beizli, Zäuerli, Baumeisterverband, Steuerrechnung, Krebsliga, Personenwagenlenker, …

Nur in Deutschland gebräuchlich:

Dezernent, Wohnungsbaugesellschaft, Existenzgründer, Flächennutzungsplan, Bahnhofsviertel, Verbraucherzentrale, Unterrichtsgarantie, Prunksitzung, Bezirksliga, Gemeindevertretung, Kreisverwaltung, umwidmen, kostenfrei, …

Im innerdeutschen Vergleich von Lemmalisten einzelner oder gebündelter Zeitungen zeigen sich sehr viel weniger Lexeme, die regional eingeschränkt gebräuchlich sind, als im nationalen Vergleich. Die Unterschiede zwischen deutschen Varietätenlemmalisten beruhen fast ausschließlich auf Eigennamen und themenspezifischem Vokabular (Berlin Großflughafen – Nürnberg Weinanbau). Durch einen Vergleich der *Berliner Morgenpost* und der *Berliner Zeitung* mit den *Nürnberger Nachrichten* und der *Nürnberger Zeitung* beispielsweise treten aber auch regionale Differenzen im Wortschatz zu Tage:

Berlin: Zwetschke, Schrippe, Weihnachtsmarkt, Baudezernent, Sonnabend, …
Nürnberg: Zwetschge, Wekkla, Christkindlesmarkt, Baureferent, Samstag, …

1.3 Verwendete Literatur

Die Internetadressen für Forschungsliteratur, Sprachressourcen, Dokumentationen und Projekte beziehen sich auf den Stand Mai 2013.

Basislemmaliste

Archer, Dawn (Hg.) (2009): What's in a Word-list? Investigating Word Frequency and Keyword Extraction. Aldershot, Ashgate.

Belica, Cyril (1994): A German Lemmatizer. Final Report MLAP93-21/WP2. Luxemburg. www.ids-mannheim.de/kl/dokumente/glemmrep.pdf

Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lüngen, Harald (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich Hermann (Hg.): Grammatik und Korpora. Dritte Internationale Konferenz. Mannheim, 22.-24.9.2009. Tübingen, Narr: 451-469.

Connexor Oy (1997-2006a): Machinese Language Model. (Confidential Information)

Connexor Oy (1997-2006b): Machinese Server. (Confidential Information)

Erb, Sabine (2005): Stichwortansetzung von Partizipien. In: Haß, Ulrike (Hg.): 91-95.

Gallmann, Peter (1991): Wort, Lexem und Lemma. In: Augst, Gerhard/Schaeder, Burkhard (Hg.): Rechtschreibwörterbücher in der Diskussion. Geschichte – Analyse – Perspektiven. Frankfurt a.M., Peter Lang: 261–280.

Haß, Ulrike (2005) (Hg.): Grundfragen der elektronischen Lexikographie. Berlin, de Gruyter.

Klosa, Annette (2005): Orthografie und morphologische Varianten. In: Haß, Ulrike (Hg.): 133-140.

Klosa, Annette (2011): Von Abbildung bis Wortelement. In: Klosa, Annette (Hg.): 157-172.

Klosa, Annette (2011) (Hg.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen, Narr.
Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica (International Annual for Lexicography) 28: 71-98.
Klosa, Annette/Schoolaert, Sabine (2011): Die lexikografische Behandlung von Eigennamen in *elexiko.* In: Klosa, Annette (Hg.): 193-211.
Knowles, Gerry/Mohd Don, Zuraidah (2004): The notion of a „lemma“: Headwords, roots and lexical sets. In: International Journal of Corpus Linguistics 9/1: 69-82. www.corpus4u.org/forum/upload/forum/2005071007011220.pdf
Kühn, Peter (1979): Der Grundwortschatz. Bestimmung und Systematisierung. Tübingen, Niemeyer.
Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010). Valletta (Malta), European Language Resources Association (ELRA): 1848-1854. www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf
Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji (Hg.): Working Papers in Corpus-based Linguistics and Language Education 3. Tokyo, University of Foreign Studies (TUFS): 53-59. http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf
Leech, Geoffrey/Rayson, Paul/Wilson, Andrew (2001): Word Frequencies in Written and Spoken English – based on the British National Corpus. Harlow, Longman.
Müller-Spitzer, Carolin (2005): Die Modellierung lexikografischer Daten und ihre Rolle im lexikografischen Prozess. In: Haß, Ulrike (Hg.): 36-54.
Pettersson, Helena/Pankow, Christiane (2006): Auswertung der Leistung von zwei frei zugänglichen POS-Taggern für die Annotation von Korpora des gesprochenen Deutsch. Göteborger Arbeitspapiere zur Sprachwissenschaft. https://gupea.ub.gu.se/handle/2077/19368
Pfeffer, J. Alan (1970): Grunddeutsch. Basic (Spoken) German Dictionary. For Everyday Usage. Englewood Cliffs (N.J.), Prentice-Hall.
Randall, L. Jones/Tschirner, Erwin (2006): A Frequency Dictionary of German. Core Vocabulary for Learners. London, Routledge.
Rosengren, Inger (1969): Wort und Worform. In: Studia Linguistica – Journal of General Linguistics 23: 103-113.
Rosengren, Inger (1972-1977): Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt, Süddeutsche Zeitung. 2 Bde. Lund, Gleerup.
Ruoff, Arno (1981): Häufigkeitswörterbuch der gesprochenen Sprache: alphabetisch, rückläufig alphabetisch und nach Häufigkeit geordnet. Tübingen, Niemeyer.
Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). www.sfs.uni-tuebingen.de/resources/stts-1999.pdf
Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf
Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT Workshop. Dublin. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf
Schnörch, Ulrich (2002): Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und lexikografischen Aufarbeitung. Tübingen, Narr.
Schnörch, Ulrich (2005): Die elexiko-Stichwortliste. In: Haß, Ulrike (Hg.): 71-90.
Storjohann, Petra (2005: Das *elexiko*-Korpus. In: Haß, Ulrike (Hg.): 55-70.
Tschirner, Erwin (2005): Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Heine, Antje/Hennig, Mathilde/Tschirner, Erwin (Hg.): Deutsch als Fremdsprache – Konturen und Perspektiven eines Fachs. München, Iudicium: 133-149. http://www.uni-leipzig.de/herder/red_tools/dl_document.php?PHPSESSID=h6727j48v2a99g3g&id=87&PHPSESSID=h6727j48v2a99g3g
Wiegand, Herbert Ernst (1983): Was ist eigentlich ein Lemma? In: Studien zur neuhochdeutschen Lexikographie 3. Hildesheim, Olms: 401-474.

Zinsmeister, Heike/Witt, Andreas/Kübler, Sandra/Hinrichs, Erhard (2008): Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In: Lüdeling, Anke/Kytö, Merja (Hg.) Corpus Linguistics. An International Handbook 1. Berlin, de Gruyter: 759-776.

Varietätenlemmalisten

Ammon, Ulrich (1995): Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietät. Berlin, de Gruyter.

Ammon, Ulrich/Bickel, Hans/Ebner, Jakob et al. (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin, de Gruyter.

Berend, Nina (2005): Regionale Gebrauchsstandards. Gibt es sie und wie lassen sie sich beschreiben. In: Eichinger, Ludwig / Kallmeyer, Werner (Hg.): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? IDS Jahrbuch 2004. Berlin, de Gruyter: 143-170.

Bickel, Hans/Hofer, Lorenz (2003): Die Lexik des schweizerischen Deutsch im GWDS. In: Wiegand, Herbert Ernst (Hg.): Untersuchungen zur kommerziellen Lexikographie der deutschen Gegenwartssprache I: Duden, das große Wörterbuch der deutschen Sprache in zehn Bänden; Print- und CD-ROM-Version. Tübingen, Niemeyer: 245-257. www.idiotikon.ch/Texte/Bickel/BickelHofer_LexikGWDS.pdf

Bickel, Hans/Landolt, Christoph (2012): Schweizerhochdeutsch. Wörterbuch der Standardsprache in der deutschen Schweiz. Mannheim, Duden.

Clyne, Michael (1992): German as a pluricentric language. In: Clyne, Michael (Hg.): Pluricentric Languages. Differing Norms in Different Nations. Berlin, de Gruyter: 117-147.

Ebner, Jakob (2009)[4]: Wie sagt man in Österreich? Wörterbuch des österreichischen Deutsch. Mannheim, Duden.

Eichhoff, Jürgen (1977-2000): Wortatlas der deutschen Umgangssprache. 4 Bde. Bern, Francke.

Eichinger, Ludwig M. (2005): Norm und regionale Variation. Zur realen Existenz nationaler Varietäten. In: Lenz, Alexandra N./Mattheier, Klaus J. (Hg.).: Varietäten – Theorie und Empirie. Frankfurt a.M., Peter Lang: 141-162.

Elspaß, Stephan/Möller, Robert: Atlas zur deutschen Alltagssprache (AdA). www.atlas-alltagssprache.de/

Goebl, Hans (2002): Dialektometrie. In: Altmann, Gabriel/Bagheri, Dariusch/Goebl, Hans/Köhler, Reinhard/Prün, Claudia: Einführung in die quantitative Lexikologie. Göttingen, Peust & Gutschmidt: 179-219.

Manning, Christopher/Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. Cambridge, MIT Press.

Meyer, Kurt (1989): Wie sagt man in der Schweiz? Wörterbuch der schweizerischen Besonderheiten. Mannheim, Duden.

Meyer, Kurt (1989): Wie sagt man in der Schweiz? Wörterbuch der schweizerischen Besonderheiten. Mannheim, Duden.

Mitzka, Walter/Schmitt, Ludwig Erich (1951-1980): Deutscher Wortatlas. 22 Bde. Giessen, Schmitz.

Retti, Gregor (1999): Austriazismen in Wörterbüchern. Zum Binnen- und Außenkodex des Österreichischen. Innsbruck, phil. Diss. http://gregor.retti.info/docs/retti1999/

Zastrow, Thomas (2011): Neue Analyse- und Visualisierungsmethoden in der Dialektometrie. Dissertation, Universität Tübingen. http://tobias-lib.uni-tuebingen.de/volltexte/2011/5919/pdf/phd.pdf

Korpora

DeReKo (2011): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2011). Mannheim. www.ids-mannheim.de/kl/projekte/korpora/archiv.html

DeReKo (2012): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2012-I (Release vom 29.02.2012). Mannheim. www.ids-mannheim.de/kl/projekte/korpora/archiv.html

Sprachressourcen, Projekte und Dokumentationen

DeReKo – Deutsches Referenzkorpus
www.ids-mannheim.de/kl/projekte/korpora/

DeReWo – Korpusbasierte Grund-/Wortformenlisten
www.ids-mannheim.de/kl/projekte/methoden/derewo.html

elexiko
www.owid.de/wb/elexiko/start.html

Machinese Phrase Tagger Demo
www.connexor.com/nlplib/?q=demo/mpt

Machinese Syntax Demo
www.connexor.com/nlplib/?q=demo/syntax

Machinese Tagset Morphologie
www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html

Machinese Tagset Syntax
www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/syntax.html

Metalemmaliste
www.metalemmaliste.de/

STTS - Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset
www.sfs.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html

TIGER-Korpus
www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

TreeTagger
www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen
www.sprache-und-genome.de/

XIP Dokumentation
http://open.xerox.com/Services/XIPParser/Pages/Using XIP

XIP Demoseite
http://open.xerox.com/Services/XIPParser/Consume/64

2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Dem Zuwendungsempfänger standen in der Projektlaufzeit vom 01.10.2008 – 31.12.2012 die folgenden Mittel zur Verfügung:

		bewilligt	Ausgaben
0812	Wiss. Angestellte	99.122 €	99.520 €
0822	Stud. Hilfskräfte	19.491 €	20.605 €
0843	Allg. Verwaltungsausgaben	1.800 €	53 €
0846	Reisekosten	3.590 €	3.794 €
		124.003 €	123.972 €

Der Umfang der tatsächlichen Ausgaben lag innerhalb des Gesamtfinanzierungsplans.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Der Verlauf der Arbeit entsprach sowohl innerhalb des geplanten Projektzeitraums als auch innerhalb des Verlängerungszeitraums mit geringen Abweichungen dem vorgesehenen Zeitplan. Die formulierten Aufgaben innerhalb des Verbundprojektes waren aufeinander abgestimmt und konnten in vollem Umfang realisiert werden. Die einzelnen Teilziele und Teilschritte wurden in den jährlichen Zwischenberichten dargelegt.

4. Voraussichtlicher Nutzen

Ein Nutzen der BLL der nhd. Standardsprache besteht in mehrfacher Hinsicht. Projektintern bildet sie das Zentrum der Metalemmaliste mit den Varietäten-Wörterbüchern, deren Wortschatz sich vom heutigen Standarddeutsch in diachroner, diatopischer oder diaphasischer Hinsicht unterscheidet. Über die Lemmaform der heutigen Standardsprache kann auf die lexikalischen Pendants einer historischen Sprachstufe, mit regionaler Ausprägung oder aus nicht formalen Kommunikationssituationen zugegriffen werden. Als Referenzliste für die Varietäten-Lemmata gibt die BLL der nhd. Standardsprache auch Auskunft über die POS-Tags und die Gebrauchshäufigkeit der Lemmata im heutigen Standarddeutsch.

Die BLL der nhd. Standardsprache steht für das interessierte Fachpublikum auch als einzelnes Produkt über das Internetportal *DeReWo – Korpusbasierte Grund-/Wortformenlisten* des IDS Mannheim zum Download bereit. Linguisten, Lexikografen, Germanisten und Philologen mit Bezug zu anderen Sprachen konstituieren die primäre Zielgruppe, aber auch in der Sprachverarbeitung und in den Informationswissenschaften kann die BLL als lexikalisches Referenzverzeichnis dienen. Das am IDS konzipierte Modell zur Erstellung der BLL der nhd. Standardsprache, das auf Sprachressourcen und Programmmodulen zu deren Analyse, Vereinheitlichung und Komparation basiert, kann zur Generierung von lexikografisch validen Lemmalisten aus anderen Korpora sowie zur Überprüfung der Güte von Lemmatisierung und POS-Tagging weiterer linguistischer Annotationswerkzeuge verwendet werden.

Die im achtmonatigen Verlängerungszeitraum gewonnenen Erkenntnisse bezüglich der Gliederung des deutschen Sprachraums durch Algorithmen, die die Lemmalisten einzelner Zeitungen verarbeiten, sowie die generierten korpusbasierten Varietätenlemmalisten können als Grundlage für die Lexikografie oder weitere dialektometrische Untersuchungen anhand der Lexik des Deutschen dienen.

5. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Auf dem Gebiet der Germanistik sind dem ZE im Projektzeitraum keine Fortschritte anderer Stellen bei der Erstellung und Auswertung korpusbasierter, frequenzsortierter Lemmalisten bekannt. Im englischsprachigen Raum wurde der Sammelband *What's in a word list? Investigating Word Frequency and Keyword Extraction* (Archer 2009) mit wissenschaftlichen Beiträgen zur Erstellung von Wortlisten für die Analyse von Frequenzen und Stichwörtern publiziert. Diese Publikation verdeutlicht das weitere Potential, das mit der Erstellung und Interpretation von korpusbasierten Wort- und Lemmalisten verbunden ist.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses

Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2012): DeReWo: Korpusbasierte Grundformenliste. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip

Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2011): DeReWo: Korpusbasierte Grundformenliste. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-250000g-2011-12-31-0.1.zip

Stadler, Heike (2013a i.E.): Die Erstellung der Basislemmaliste der nhd. Standardsprache aus mehrfach linguistisch annotierten Korpora. Sammelband zum Abschlussworkshop des Wechselwirkungs-Projektes: Varianz und Vielfalt interdisziplinär: Wörter und Strukturen. Darmstadt, 29.11.-30.11.2012. http://www.sprache-und-genome.de (pdf)

Stadler, Heike (2013b): Automatische Lemmatisierung und POS-Tagging in der Praxis. Dokumentation der im Wechselwirkungs-Projekt gemachten Erfahrungen mit den linguistischen Annotations-werkzeugen glemm, Machinese Phrase Tagger, Machinese Syntax, TreeTagger, Xerox XIP. IDS-interne Dokumentation.

Stadler, Heike (2012): Technische Dokumentation zur Erstellung der Basislemmaliste der nhd. Standardsprache aus mehrfach linguistisch annotierten Korpora. IDS-interne Dokumentation.

Wegstein, Werner/Stadler, Heike (2012): Why and how to encode word structures and word-formation formula of a word family dictionary? A proposal based on <etym>. TEI-Conference TEI and the C(r|l)o(w|u)d, College Station, Texas, 7.11.-10.11.2012. http://www.sprache-und-genome.de (pdf)

Mannheim, 24. Juni 2013

Prof. Dr. Dr. h.c. mult. Ludwig M. Eichinger

Berichtsblatt Förderkennzeichen 01UA0815B

1. ISBN oder ISSN geplant	2. Berichtsart Schlussbericht

3a Titel des Berichts
Die Basislemmaliste der neuhochdeutschen Standardsprache

3b Titel der Publikation
Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2012): DeReWo: Korpusbasierte Grundformenliste.
http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip

3c Titel der Publikation
Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2011): DeReWo: Korpusbasierte Grundformenliste.
http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-250000g-2011-12-31-0.1.zip

3d Titel der Publikation
Stadler, Heike (2013 i.E.): Die Erstellung der Basislemmaliste der nhd. Standardsprache aus mehrfach linguistisch annotierten Korpora. Sammelband zum Abschlussworkshop des Wechselwirkungs-Projektes: Varianz und Vielfalt interdisziplinär: Wörter und Strukturen. Darmstadt, 29.11.-30.11.2012.

3e Titel der Publikation
Wegstein, Werner/Stadler, Heike (2012): Why and how to encode word structures and word-formation formula of a word family dictionary? A proposal based on <etym>. TEI-Conference TEI and the C(r|l)o(w|u)d, College Station, Texas, 7.11.-10.11.2012.

4. Autor(en) [Name(n), Vorname(n)] Stadler, Heike	5. Abschlussdatum des Vorhabens 31.12.2012
	6. Veröffentlichungsdatum geplant
	7. Form der Publikation TIB
8. Durchführende Institution(en) (Name, Adresse) Institut für Deutsche Sprache R5, 6-13 68161 Mannheim	9. Ber. Nr. Durchführende Institution -
	10. Förderkennzeichen 01UA0815B
	11. Seitenzahl 37
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 73
	14. Tabellen 3
	15. Abbildungen 8

16. Zusätzliche Angaben
-

17. Vorgelegt bei (Titel, Ort, Datum)
-

18. Kurzfassung
Die Bestimmung von Varianz und Varietät in der Sprache wird im Wechselwirkungs-Projekt auf der lexikalisch morphologischen Ebene über den Vergleich von Varietäten-Wörterbüchern mit der Basislemmaliste (BLL) der nhd. Standardsprache realisiert. Die BLL konstituiert als Tertium Comparationis das Zentrum in einem diachronen und dialektalen Vergleich. Als Grundlage für die Erstellung der BLL dient das Deutsche Referenzkorpus DeReKo, dessen Wortformen mit den linguistischen Annotationen mehrerer sprachverarbeitender Werkzeuge angereichert sind. In den automatischen Verarbeitungsprozess der linguistischen Annotationen wird das Wissen eines Wörterbuchs integriert, so dass die BLL mit lexikografischen Konventionen korreliert. Die frequenzsortierte BLL der nhd. Standardsprache enthält mehr als 326.000 Lemmata mit Part-of-Speech-Tags und Häufigkeitsangaben. Eine korpusbasierte Liste mit 60.000 Eigennamen wird zusätzlich zur Verfügung gestellt. Über die Webseiten der Deutschen Referenzwortliste DeReWo am IDS Mannheim steht die BLL anderen Wissenschaftlern und der interessierten Öffentlichkeit frei zur Verfügung und kann als Referenzlemmaliste der nhd. Standardsprache in der Linguistik, Informatik und Computerlinguistik Verwendung finden.

19. Schlagwörter
Korpuslinguistik, maschinelle Lexikografie, automatische linguistische Annotation

20. Verlag -	21. Preis -

Document Control Sheet 01UA0815B

1. ISBN or ISSN under preparation	2. type of document final report
3a title of final report The Compilation of a German Lemma List from Large Corpora	
3b title of publication Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2012): DeReWo: Korpusbasierte Grundformenliste. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip	
3c title of publication Perkuhn, Rainer/Stadler, Heike/Belica, Cyril/Kupietz, Marc (2011): DeReWo: Korpusbasierte Grundformenliste. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-250000g-2011-12-31-0.1.zip	
3d title of publication Stadler, Heike (2013 i.E.): Die Erstellung der Basislemmaliste der nhd. Standardsprache aus mehrfach linguistisch annotierten Korpora. Sammelband zum Abschlussworkshop des Wechselwirkungs-Projektes: Varianz und Vielfalt interdisziplinär: Wörter und Strukturen. Darmstadt, 29.11.-30.11.2012.	
3e title of publication Wegstein, Werner/Stadler, Heike (2012): Why and how to encode word structures and word-formation formula of a word family dictionary? A proposal based on <etym>. TEI-Conference TEI and the C(r|l)o(w|u)d, College Station, Texas, 7.11.-10.11.2012.	
4. author(s) (family name, first name(s)) Stadler, Heike	5. end of project 12-31-2012
	6. publication date under preparation
	7. form of publication TIB
8. performing organization(s) (name, address) Institut für Deutsche Sprache R5, 6-13 68161 Mannheim	9. originator's report no. -
	10. reference no. 01UA0815B
	11. no. of pages 37
12. sponsoring agency (name, address) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. no. of references 73
	14. no. of tables 3
	15. no. of figures 8
16. supplementary notes -	
17. presented at (title, place, date) -	
18. abstract In the project *Interaction between linguistic and bioinformatic processes, methods and algorithms – modelling and visualizing variance in language and genomes*, the identification of variance and variety in the German language is determined on a lexical-morphological level, comparing variety dictionaries with a lemma list of the New High German Standard Language. The lemma list of New High German serves as a Tertium Comparationis and forms the basis in a diachronic and dialectal space. The lemma list is compiled from the German Reference Corpus DeReKo, which was enriched with linguistic annotations by several language processing tools. The programming of a modular toolbox for the automated processing and comparison of the linguistic corpus annotations combines alignment algorithms with digitised lexicological knowledge so that the automatically generated corpus-based lemma list observes lexicographic conventions. The frequency-sorted lemma list of New High German contains more than 326 000 lemmata with part-of-speech and frequency information. In addition, a corpus-based list with more than 68 000 proper names is provided. The lemma list is available for scientists and the interested general public through the websites of the German Reference Wordlist DeReWo at IDS Mannheim.	
19. keywords Corpus linguistics, electronical lexikography, automatic linguistic annotation	
20. publisher -	21. price -