Benutzerdokumentation

Technical Report IDS-KL-2013-02

zum Produkt

Korpusbasierte Wortgrundformenliste

DeReWo
v-ww-bll-320000g-2012-12-31-1.0

Institut für Deutsche Sprache, Mannheim
Januar 2013

Inhaltsverzeichnis

Vorwort

Dieses Dokument orientiert sich in der Struktur an den allgemeinen Bemerkungen zu der Reihe DeReWo (DeReWo 2013), die der Leser in der aktuellen Fassung zur Kenntnis genommen haben sollte. An dieser Stelle werden die dort skizzierten Problembereiche nicht erneut aufgerollt, sondern es werden nur die konkreten Entscheidungen im Rahmen des vorliegenden Produkts dokumentiert.
Die vorliegende Lemmaliste wurde im IDS-Teilprojekt des Verbundvorhabens „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen“ (http://www.ids-mannheim.de/kl/projekte/ww/) von Heike Stadler erarbeitet. Sie ist eine Weiterentwicklung des Produkts DeReWo v-ww-bll-250000g-2011-12-31-0.1 (DeReWo 2011). Eine ausführliche, projektbezogene Dokumentation zur Erstellung der Liste kann bei Bedarf über den u.g. Kontakt angefordert werden.

Download

Das Original dieser DeReWo-Wortliste kann unter http://www.ids-mannheim.de/derewo zusammen mit der Dokumentation in der jeweils aktuellen Version abgerufen werden.
Bitte beachten Sie die Lizenzbestimmungen am Ende dieses Dokuments.

1 Grundsätzliches zu DeReWo v-ww-bll-320000g-2012-12-31-1.0

1.1 Was ist DeReWo v-ww-bll-320000g-2012-12-31-1.0?

DeReWo v-ww-bll-320000g-2012-12-31-1.0 ist die Lemmastrecke eines fiktiven Wörterbuchs des öffentlichen Schriftsprachgebrauchs der letzten 30 Jahre im Umfang von ca. 320.000 Lemmata zusammengestellt auf der Grundlage herkömmlicher lexikographischer Kriterien mit besonderer Berücksichtigung der Gebrauchshäufigkeit in Form von Korpusfrequenz.

1.2 Wie wurde DeReWo v-ww-bll-320000g-2012-12-31-1.0 erstellt?

DeReWo v-ww-bll-320000g-2012-12-31-1.0 wurde in einer Kombination von automatischen, semi-automatischen und manuellen Verfahren erstellt, je nachdem, welche Vorgehensweise zur Lösung welcher partiellen Problemstellung sinnvoll bzw. erforderlich war.

2 Methodik im Einzelnen

Die Methodik ist ausführlich in der Dokumentation des Projekts Wechselwirkungen beschrieben. Die Vorgehensweise bei der Erstellung dieser Lemmaliste weicht insofern von älteren DeReWo-Studien ab, als dass verstärkt versucht wird, die Informationen mehrerer, größtenteils automatisch erstellter Ressourcen aufeinander zu beziehen. Übereinstimmungen werden als Evidenz für die Zuverlässigkeit der Information gedeutet, Nicht-Übereinstimmungen führen zu eingehenderen Analysen. Zusätzliche Informationen manueller Auswertungen werden dann miteinbezogen, um über Abbildungen der Angaben aufeinander den Begriff der Übereinstimmung ggf. zu erweitern.

2.1 Ressourcen

2.1.1 Korpusbasiertheit

Der Grundformenliste liegen die Korpora des DeReKo-Archivs Stand Mitte 2012 (DeReKo 2012) zugrunde.

2.1.2 Vergleichslemmastrecke

Zum Abgleich wurde eine weitere Ressource zusammengestellt, die initial aus den Stichwortstrecken verschiedener Wörterbücher, u.a. auch des Projekts *elexiko* (elexiko 2011) bestand. Diese Lemmastrecke wurde um weitere, zuvor fehlende Einträge ergänzt, die beim Abgleich mit den anderen Ressourcen aufgefallen waren, manuell jedoch als relevante Stichwörter bewertet wurden.

2.1.3 Wortklassenangaben

Die Wortklassenangaben, die verschiedene Werkzeuge bereitstellen (*part-of-speech tags*), wurden zur Erkennung von mehrdeutigen Einträgen (z.B. *'sein'* als Verb oder Pronomen) eingesetzt und zur Kennzeichnung dieser Fälle mit angegeben. Eine Aussage über die Qualität dieser automatisch ermittelten Angaben kann an dieser Stelle nicht gemacht werden.

2.2 Wortformenlisten

Für DeReWo v-ww-bll-320000g-2012-12-31-1.0 wurde als eine Ressource eine Wortformenliste des Korpusarchivs mit hausinternen Werkzeugen erstellt.
Als weitere Ressourcen wurden diverse Grundformenlisten unmittelbar aus den entsprechenden Korpora ermittelt, die mit verschiedenen Werkzeugen (*TreeTagger* (Schmid 1994, 1995), Machinese Phrase Tagger/Syntax[1]) bearbeitet worden waren.

2.3 Grundformenlisten

Die erste Ressource, die Wortformenliste, wurde mithilfe des Werkzeugs *glemm* (Belica 1994) lemmatisiert, die weiteren Ressourcen lagen bereits als Grundformenlisten vor. Alle spezifischen Lemmalisten enthielten die Angabe der Grundformen, ihrer automatischen bestimmten Wortklasse, sowie ihrer Gebrauchshäufigkeit. Als allgemeine Vorgehensweise wurden die in dieser Form zur Verfügung stehenden Lemmalisten mit der Vergleichslemmastrecke abgeglichen. Ergaben sich Übereinstimmungen von dieser mit mindestens zwei weiteren Ressourcen, wurde ein Eintrag mit dem Mittelwert der dort gebuchten Frequenzen festgehalten. Nicht-Übereinstimmungen zwischen den Lemmalisten (hinsichtlich Grundformangabe oder Wortklassenzugehörigkeit) bzw. mit der Vergleichslemmastrecke führten zu einer Analyse hinsichtlich Lemmazugehörigkeit und Nennform und ggf. zu einer Neueinordnung und entsprechender Verrechnung, ggf. zu einer Erweiterung der Vergleichslemmastrecke.

2.3.1 Groß-/Kleinschreibung

Die Unterscheidung der Schreibung wird gehandhabt wie bei Erstellung der Ressourcen bzw. beim Einsatz der Werkzeuge festgelegt. Davon abweichend wird eine Schreibweise nicht der

[1] s. http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html und http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/syntax.html
vgl. auch http://www.connexor.com/nlplib/

anderen zugeordnet, auch wenn sie nicht in der Vergleichslemmastrecke belegt ist, falls ihre Häufigkeit die der anderen Schreibweise übersteigt (etwa bei substantivierten Verben). In diesen Fällen wird ein eigenständiges Lemma angelegt.

2.3.2 Trennzeichen/Bindestrich

Die Unterscheidung der Schreibung wird gehandhabt wie bei Erstellung der Ressourcen bzw. beim Einsatz der Werkzeuge festgelegt.

2.3.3 Diskontinuierliche Konstituenten, Präverbfügungen (abtrennbare/abgetrennte Präfixe)

Diese Fragestellung wird für die aktuelle Version dahingehend zurückgestellt, dass keine Rekonstruktion der richtigen zahlenmäßigen Verhältnisse angestrebt wurde. Sofern von den Werkzeugen erkannt, wird aber die Kennzeichnung einer Form als Präverb (VRZ = Verbzusatz) ggf. mit in die Liste übernommen. Soweit erkennbar werden unvollständige Bestandteile elliptischer Formulierungen (Im- und Export) für die weitere Verarbeitung nicht berücksichtigt.

2.3.4 Varianten/Varietäten (regional, diachron, Rechtschreibreform, Getrennt-/Zusammenschreibung)

Für orthografische Varianten werden Informationen, die das Projekt *elexiko* (elexiko 2011) zur Verfügung gestellt hat, zu Rate gezogen. Diese werden zur vorgeschlagenen präferierten Nennform zusammengefasst.

2.3.5 Neubildungen/Neologismen

Dieser Punkt wird im Rahmen der allgemeinen Vorgehensweise (s. Abschn. 2.3) behandelt.

2.3.6 Adjektivisch gebrauchte Partizipien

Ein Partizip wird dann in die Lemmastrecke übernommen, wenn es auch in der Vergleichslemmastrecke belegt ist, oder wenn seine Formen häufiger sind als die restlichen des Verbparadigmas; anderenfalls werden sie der Infinitivangabe des Verbs zugerechnet.

2.3.7 Nennung der Grundform

Die Lemmata werden in der Form genannt wie sie von den Tools übereinstimmend vorgeschlagen werden – außer wenn *elexiko* Schreibvarianten vermerkt hat (s. Varianten 2.3.4). In Ausnahmefällen bzw. bei Nicht-Übereinstimmungen wird eine geeignete Form ausgewählt bzw. eine künstliche Form gebildet (z.B. *'der,die,das'* oder *'dies(e,er,es)'*).

2.3.7.1 Movierung

Die Lemmata werden im Standardfall in der Form übernommen wie sie von den Tools vorgeschlagen werden – sofern in der Vergleichslemmastrecke vorhanden. Da diese sehr viele movierte Formen enthält, gleichzeitig die Werkzeuge Movierungen aber sehr uneinheitlich handhaben, kann dies zu Ungenauigkeiten bei der Lemma- und der Häufigkeitsangabe führen.

2.3.7.2 Reflexive Verben

Die Bestandteile reflexiver Verben werden als zwei getrennte Lemmata behandelt.

2.3.7.3 Pronomen

Pronomen werden in der Vielfalt übernommen, wie sie von den Werkzeugen vorgeschlagen werden und in der Vergleichslemmastrecke belegt sind.

2.3.8 Abgleich mit Wörterbuchlemmastrecken

Die Rohlisten werden mit der Wörterbuchvergleichslemmastrecke abgeglichen. Die in der Lemmastrecke nicht belegten Kandidaten werden hinsichtlich Lemmazugehörigkeit und Nennform überprüft und ggf. neu eingeordnet und verrechnet. Von den Werkzeugen entsprechend markierte oder beim Abgleich nicht belegte Kandidaten der Lemmalisten werden speziell im Hinblick auf die u.g. Kriterien der Relevanz-Sonderfälle manuell eingeordnet.

2.4 Relevanz-Sonderfälle

2.4.1 Fremdwörter, Anglizismen

Diese Eigenschaft wird aufgrund Taggerinformation bzw. manueller Auswertung gekennzeichnet, führt jedoch nicht zu einer Herausfilterung – sofern der Wörterbuchabgleich oder eine hohe manuelle Relevanzeinschätzung dem entgegenstehen.

2.4.2 Eigennamen

Eigennamen werden aufgrund Taggerinformation bzw. manueller Auswertung gekennzeichnet und herausgefiltert, auch wenn sie in der Vergleichslemmastrecke verzeichnet sind – außer, wenn sie nicht automatisch oder über Abgleiche als solche erkannt werden können.

2.4.3 Wortreihen

Dieser Punkt erscheint durch die angestrebte Größenordnung wenig relevant, wird jedoch für Einzelfälle bei der händischen Auswertung berücksichtigt.

2.4.4 Kurzwörter

Dieser Punkt wird im Rahmen der allgemeinen Vorgehensweise (s. Abschn. 2.3) behandelt. Sie werden aufgenommen, sofern sie in der Vergleichslemmastrecke enthalten sind.

2.4.5 Akronyme, Einzelbuchstaben und Kürzel

Dieser Punkt wird im Rahmen der allgemeinen Vorgehensweise (s. Abschn. 2.3) behandelt. Akronyme und Kürzel werden nicht aufgenommen, da sie in der Vergleichslemmastrecke nicht enthalten sind und auch von den Werkzeugen sehr uneinheitlich behandelt werden. Einzelbuchstaben werden herausgefiltert.

2.4.6 Unselbstständige Morpheme

Dieser Punkt wird im Rahmen der allgemeinen Vorgehensweise (s. Abschn. 2.3) behandelt. Sie werden nicht aufgenommen, da sie in der Vergleichslemmastrecke nicht enthalten sind.

2.4.7 Verschmelzungen (Amalgamierung) (*ans*, *zum*, *zur*, *fürs*)

Dieser Punkt wird im Rahmen der allgemeinen Vorgehensweise (s. Abschn. 2.3) behandelt. Im Gegensatz zu früheren Listen werden „PREP+DET“-Wortformen keiner anderslautenden Grundform zugeordnet, sondern selbst als Lemma („PREP“) angesetzt.

2.5 Häufigkeitsklassen

Die Häufigkeit einer Grundform in absoluten Zahlen anzugeben ist wenig sinnvoll. Der Betrachter verbindet damit eine Genauigkeit und eine Zuverlässigkeit der Aussage, die nicht gegeben ist. Aufgrund der Zusammensetzung der Datengrundlage können sich Verzerrungen bei den Grundformfrequenzen ergeben. Als relativ stabil und aussagekräftig – gerade auch beim Vergleich unterschiedlich großer Datenbestände – hat sich erwiesen, Häufigkeiten in Form von Häufigkeitsklassen anzugeben[2]. Dabei hat eine Grundform die Häufigkeitsklasse N, wenn die häufigste Form etwa 2^N-mal häufiger vorkommt als diese Form. Für die Grundformenliste ist der Eintrag mit der höchsten Frequenz *'der,die,das'* mit f(*'der,die,das'*)= 490.092.568, d.h.

$$N = hk(grundform) := \lfloor \log_2(f('der,die,das')/f(grundform)) + 0{,}5 \rfloor$$

also $f(grundform) \approx f('der,die,das')/2^N$.

Bsp.

N =	0	1	2	3	4	5		10		29
2^N =	2^0	2^1	2^2	2^3	2^4	2^5	…	2^{10}	…	2^{29}
2^N =	1	2	4	8	16	32		1.024		536.870.912
Bsp.	*der, die, das*	-	*und*	*mit*	*als*	*Jahr*		*greifen*		*Unleidigkeit*

D.h. *'der, die, das'* ist etwa vier Mal so häufig wie *'und'*, etwa acht Mal so häufig wie *'mit'* und etwa 536.870.912 Mal[3] so häufig wie *'Unleidigkeit'*.

In ihrem Originalzustand ist die Liste auch innerhalb der Häufigkeitsklassen nach der absoluten Häufigkeit sortiert!

2.6 Qualitätskontrolle

Eine Qualitätskontrolle erfolgte stets begleitend bei der Analyse der Sonderfälle.

3 Dateiformat

Die Grundformenliste ist als Datei mit dem Namen

derewo v-ww-bll-320000g-2012-12-31-1.0.txt

dem Archiv beigelegt. Sie ist im Zeichensatz ISO-8859-15 kodiert.

Nach einem 43-zeiligen Header, der die Hinweise auf die Lizenzbedingungen enthält und der mit „# " am Zeilenanfang als Kommentar gekennzeichnet ist, sind die Einträge der Grundformenliste zeilenweise dreispaltig angegeben: Das erste Feld enthält die Grundform, davon mit einem Leerzeichen abgetrennt ist im zweiten Feld deren Häufigkeitsklasse angegeben. In der dritten Spalte ist nur in den Fällen eine Angabe zur Wortklasse vermerkt, wenn mehrere gleichlautende Formen mit verschiedenen Wortklassen in der Liste enthalten sind. Im Originalzustand ist die Liste innerhalb der Häufigkeitsklassen nach der absoluten Häufigkeit sortiert.

---

[2] unter der Voraussetzung, dass jeweils dieselbe Definition/Operationalisierung für das Maß zugrunde gelegt ist.

[3] Zur Beachtung: HK gleich Anzahl der Verdopplungsschritte bis mindestens Anzahl des häufigsten Lemmas erreicht ist; deshalb kann diese Zahl auch größer als dessen Häufigkeit sein.

Referenzen

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

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

DeReWo (2013): *Korpusbasierte Wortlisten DeReWo*, Allgemeine Anmerkungen, Technical Report IDS-KL-2013-01, http://www.ids-mannheim.de/derewo, Stand: 2013.

DeReWo (2011): *Korpusbasierte Wortgrundformenliste DeReWo*, v-ww-bll-250000g-2011-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2011.

elexiko (2011): Online-Wörterbuch des Instituts für Deutsche Sprache zur deutschen Gegenwartssprache. http://www.owid.de/wb/elexiko/start.html

Helmut Schmid (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Lizenzbestimmungen

(zu zitieren als:)

> Korpusbasierte Wortgrundformenliste DeReWo, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation,
> http://www.ids-mannheim.de/derewo,
> © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2013.

Die Wortgrundformenliste, die Dokumentation und die allgemeinen Anmerkungen bilden eine Einheit. Diese Lizenzbestimmung darf aus keinem der Dokumente entfernt werden.

Kontakt

Falls Sie speziellere Informationen benötigen, als dieses Werk bereithält, oder Sie dieses Werk über die eingeräumten Rechte hinaus nutzen möchten, wenden Sie sich bitte an derewo@ids-mannheim.de.

Bei Veröffentlichung auf diesem Werk aufbauender Forschungsergebnisse bitten wir um eine kollegiale Mitteilung an derewo@ids-mannheim.de.