Benutzerdokumentation

Technical Report IDS-KL-2011-02

zum Produkt

Korpusbasierte Wortgrundformenliste

DeReWo
v-ww-bll-250000g-2011-12-31-0.1

Institut für Deutsche Sprache, Mannheim
Dezember 2011

Inhaltsverzeichnis

Vorwort

Dieses Dokument orientiert sich in der Struktur an den allgemeinen Bemerkungen zu der Reihe DeReWo (DeReWo 2009a), 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. Diese Version wurde bereits verbundintern zur Verfügung gestellt, stellt aber nur einen Zwischenstand der Arbeiten zur Erstellungsmethodik von Lemmalisten dar. Insofern sollte sie als β-Version betrachtet werden, die insbesondere eventuell unvollständig ist und daher für bestimmte Zwecke nur bedingt verwendet werden kann. Zum Ende des Projekts (April 2012) wird eine finale Version einer Basislemmaliste veröffentlicht, die auch über die DeReWo-Reihe zugänglich gemacht werden wird.

Download

Das Original dieser DeReWo-Wortliste kann unter http://www.ids-mannheim.de/kl/derewo/ zusammen mit der Dokumentation in der jeweils aktuellen Version abgerufen werden.

1 Grundsätzliches zu DeReWo v-ww-bll-250000g-2011-12-31-0.1

1.1 Was ist DeReWo v-ww-bll-250000g-2011-12-31-0.1?

DeReWo v-ww-bll-250000g-2011-12-31-0.1 ist die Lemmastrecke eines fiktiven Wörterbuchs des öffentlichen Schriftsprachgebrauchs der letzten 30 Jahre im Umfang von 250.000 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-250000g-2011-12-31-0.1 erstellt?

DeReWo v-ww-bll-250000g-2011-12-31-0.1 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 wird ausführlich in der Dokumentation des Projekts Wechselwirkungen beschrieben werden.

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 2011 (DeReKo 2011) zugrunde.

2.1.2 Wörterbuchvergleichslemmastrecke

Zum Abgleich wurde eine weitere Ressource zusammengestellt, die aus den Stichwortstrecken verschiedener Wörterbücher, u.a. auch des Projekts *elexiko* (elexiko 2011) besteht.

2.1.3 Wortklassenangaben

Die Wortklassenangabe, die die 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 auch in der veröffentlichten Version 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-250000g-2011-12-31-0.1 wurde – als eine von mehreren Ressourcen – auf eine interne Wortformenliste, die in der Studie (DeReWo 2009b) entstanden ist, zurückgegriffen.

Als zweite Ressource wurde eine Grundformenliste unmittelbar aus den entsprechenden aktuellen Korpora ermittelt, die mit dem Werkzeug *TreeTagger* (Schmid 1994, 1995) bearbeitet worden waren.

2.3 Grundformenlisten

Die erste Ressource, die Wortformenliste, wurde mithilfe des Werkzeugs *glemm* (Belica 1994) lemmatisiert, die zweite Ressource lag bereits als Grundformenliste vor.

Die dadurch zur Verfügung stehenden zwei Lemmalisten wurden mit der WB-Lemmastrecke abgeglichen. Ergaben sich Übereinstimmungen bei allen drei Ressourcen, wurde ein Eintrag mit dem Mittelwert der Frequenzen festgehalten. Nicht-Übereinstimmungen zwischen den Lemmastrecken bzw. mit der mit WB-Lemmastrecke führten zu einer Analyse hinsichtlich

Lemmazugehörigkeit und Nennform und ggf. zu einer Neueinordnung und entsprechender Verrechnung (s. auch 2.3.8).

2.3.1 Groß-/Kleinschreibung

Die Unterscheidung der Schreibung wurde gehandhabt wie bei Erstellung der Ressourcen bzw. beim Einsatz der Werkzeuge festgelegt. Davon abweichend wurde eine Schreibweise nicht der anderen zugeordnet, auch wenn sie nicht in WB-Strecke belegt war, falls ihre Häufigkeit die der anderen Schreibweise überstieg (etwa bei substantivierten Verben). In diesen Fällen wurde ein eigenständiges Lemma angelegt.

2.3.2 Trennzeichen/Bindestrich

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

2.3.3 Diskontinuierliche Konstituenten, Präverbfügungen (abtrennbare/abgetrennte Präfixe)

Diese Fragestellung wurde für die aktuelle Version dahingehend zurückgestellt, dass keine Rekonstruktion der richtigen zahlenmäßigen Verhältnisse angestrebt wurde. Sofern von den Werkzeugen erkannt, wurde aber die Kennzeichnung einer Form als Präverb (VRZ = Verbzusatz) mit in die Liste übernommen.

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

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

2.3.5 Neubildungen/Neologismen

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.3.6 Adjektivisch gebrauchte Partizipien

Partizipien wurden dann in die Lemmastrecke übernommen, wenn sie auch in der WB-Strecke belegt waren; anderenfalls wurden sie der Infinitivangabe des Verbs zugerechnet.

2.3.7 Nennung der Grundform

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

2.3.7.1 Movierung

s.o.

2.3.7.2 Reflexive Verben

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.3.8 Abgleich mit Wörterbuchlemmastrecken

Die Rohlisten wurden mit einer Wörterbuchvergleichslemmastrecke abgeglichen. Die in der Strecke nicht belegten Kandidaten wurden hinsichtlich Lemmazugehörigkeit und Nennform überprüft und ggf. neu eingeordnet und verrechnet.
Vom TreeTagger entsprechend markierte oder beim Abgleich nicht belegte Kandidaten wurden speziell im Hinblick auf die u.g. Kriterien der Relevanz-Sonderfälle manuell eingeordnet.

2.4 Relevanz-Sonderfälle

2.4.1 Fremdwörter, Anglizismen

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

2.4.2 Eigennamen

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

2.4.3 Wortreihen

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.4.4 Kurzwörter

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.4.5 Akronyme, Einzelbuchstaben und Kürzel

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.4.6 unselbstständige Morpheme

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt.

2.4.7 Verschmelzungen (Amalgamierung) (*ans*, *zum*, *zur*, *fürs*, *fortan*, *infolge*, *aufgrund*, *zuhause*)

Diese Fragestellung wurde für die aktuelle Version nicht explizit behandelt. Im Gegensatz zu früheren Listen wurden diese Wortformen keiner anders lautenden Grundform zugeordnet, sondern selbst als Lemma 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. 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'*)= 373.738.420, 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		28
2^N =	2^0	2^1	2^2	2^3	2^4	2^5	…	2^{10}	…	2^{28}
2^N =	1	2	4	8	16	32		1.024		268.435.456
Bsp.	*der, die, das*	-	*und*	*mit*	*als*	*Jahr*		*greifen*		*Zielbewusstheit*

D.h. *'der, die, das'* ist etwa vier Mal so häufig wie *'und'*, etwa acht Mal so häufig wie *'mit'* und etwa 268.435.456 Mal so häufig wie *'Zielbewusstheit'*.

In der veröffentlichten Form 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-250000g-2011-12-31-0.1.txt

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

Nach einem 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. In der veröffentlichten Form ist die Liste auch innerhalb der Häufigkeitsklassen nach der absoluten Häufigkeit sortiert.

Referenzen

Belica, Cyril (1994). A German Lemmatizer. Final Report MLAP93-21/WP2. Luxemburg.

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

DeReWo (2007): Korpusbasierte Wortgrundformenliste DeReWo, v-30000g-2007-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2007.

DeReWo (2009a): Korpusbasierte Wortlisten DeReWo, Allgemeine Anmerkungen, http://www.ids-mannheim.de/kl/derewo/, Stand: 2009.

DeReWo (2009b): Korpusbasierte Wortgrundformenliste DeReWo, v-40000g-2009-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2009.

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-250000g-2011-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/,
> © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2011.

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.