Benutzerdokumentation

Technical Report IDS-KL-2009-02

zum Produkt

Korpusbasierte Wortformenliste

DeReWo
v-100000t-2009-04-30-0.1

Institut für Deutsche Sprache, Mannheim
April 2009

Inhaltsverzeichnis

Vorwort

Dieses Dokument orientiert sich in der Struktur an den allgemeinen Bemerkungen zu der Reihe DeReWo (DeReWo 2009), 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.

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-100000t-2009-04-30-0.1

1.1 Was ist DeReWo v-100000t-2009-04-30-0.1?

DeReWo v-100000t-2009-04-30-0.1ist eine bestimmte Sicht auf den Wortformenbestand des Deutschen Referenzkorpus DeReKo zum Stand April 2009 (DeReKo 2009). Die Liste umfasst die 100.000 häufigsten Wortformen, die in möglichst vielen verschiedenen Quellen vorkommen.

1.2 Wie wurde DeReWo v-100000t-2009-04-30-0.1 erstellt?

DeReWo v-100000t-2009-04-30-0.1 wurde weitestgehend vollautomatisch basierend auf dem Deutschen Referenzkorpus DeReKo erstellt.

2 Methodik im Einzelnen

2.1 Korpusbasiertheit

Der Wortformenliste liegen alle Korpora des DeReKo-Archivs Stand April 2009 (DeReKo 2009) zugrunde.

2.2 Wortformenlisten

Für DeReWo v-100000t-2009-04-30-0.1 gehen wir von folgenden Annahmen aus bzw. legen wir folgende Vereinbarungen fest: Wortbestandteile sind die alphabetischen Zeichen a-z, A-Z inkl. der Umlaute ä, ö, ü, Ä, Ö und Ü und diakritischer Varianten[1], sowie das ß. Die Worttrenner sind alle anderen Zeichen, insbesondere Satzzeichen, Leerzeichen und Zeilenumbrüche (außer bei Worttrennung am Zeilenende). Trennstriche beim Zeilenumbruch wurden aufgelöst (d.h. die Bestandteile auf den verschiedenen Zeilen ohne Trennstrich zusammengezogen, Spezialfall kk wird wieder zu ck: *Zuk-ker* zu *Zucker*), der Punkt wird als Trennzeichen interpretiert. Der Bindestrich ist gerade in Zeiten von Internetadressen und den verschiedenen Phasen der Rechtschreibreform sehr schwierig einheitlich zu handhaben. Er wird in unserem Fall nicht als

[1] æ Æ ø Ø á à â ã À À Â Ã é è ê È È Ê í ì î Í Ì Î ó ò ô õ Ó Ò Ô Õ ñ Ñ ç Ç

Wortbestandteil betrachtet. Zwischen Groß- und Kleinschreibung wird auf Wortformenebene konsequent unterschieden.

2.2.1 Groß-/Kleinschreibung

Bei der Tokenisierung wurde zwischen Groß- und Kleinschreibung unterschieden.

2.2.2 Trennzeichen/Bindestrich

Der Trennstrich ist bei der Tokenisierung – soweit möglich – aufgelöst worden. Der Bindestrich wird in unserem konkreten Fall nicht als Bestandteil einer Wortform gedeutet.

2.2.3 Diskontinuierliche Konstituenten, Präverbfügungen (abtrennbare/abgetrennte Präfixe)

Für diese Wortformenliste wurden die Bestandteile diskontinuierlicher Konstituenten (z.B. Präverbfügungen und Präfixe) getrennt gezählt, es wurde keine Zusammenführung vorgenommen.

2.2.4 „Ausgewogenheit", Streuung

Die Effekte einer „unausgewogenen Verteilung" bestimmter Phänomene wurden dadurch ein wenig gedämpft, dass – um auf die gewünschte Zielgröße zu kommen – nur die Wortformen berücksichtigt werden, die in mindestens 179 von 259 verwendeten Quellen beobachtet wurden. Je größer der Umfang der Zielwortformenliste jedoch ist, desto kritischer ist der Zusammenhang zwischen Häufigkeit und Streuung zu hinterfragen.

2.2.5 Rechtschreibreform, insbes. Getrennt-/Zusammenschreibung

Verschiedene Faktoren können die Schreibweisen in den Quellen je nach Stadium und Akzeptanz der Rechtschreibreform beeinflussen. Eine stichprobenartige Überprüfung deutet zumindest für den oberen Bereich der sehr häufigen Wortformen an, dass nicht für jede der verschiedenen Schreibweisen eine ausreichende Streuung in der Gesamtdatengrundlage vorliegt. Im Fall des Wortes *daß/dass* lassen sich „alte Schreibweisen" noch, „neue Schreibweisen" schon in ausreichend vielen Dokumenten nachweisen (caveat: auch früher und in anderen Sprachräumen evtl. schon als Nebenschreibweisen verzeichnet?).

Position in Liste	*Wortform*	*absolute Häufigkeit*	*Dokumenthäufigkeit*
53:	*dass*	7090644	230
64:	*daß*	5086989	250

In anderen Fällen wie z.B. bei *musste* oder *mussten* (immerhin auf Position 511 bzw. 1076 aufgrund der Gesamthäufigkeit) reicht die 179/259-Streuung über die Gesamtdatengrundlage nicht aus, um unter den ersten 100.000 berücksichtigt zu werden. Um diesem Einschnitt des Wortbestandes der deutschen Sprache gerecht zu werden, wurde für die Wörtern, die aufgrund der Rechtschreibreform in verschiedenen Schreibweisen erscheinen, ein weicheres Streuungsmaß angewandt.

2.2.5.1 Vorgehen beim weicheren Streuungsmaß

Die Korpora wurden aufgrund der Vorkommens des Wortes *dass/daß* in drei Kategorien aufgeteilt: die Korpora, in denen eine Schreibweise deutlich überwiegt (hier konkret: sich um mindestens zwei Häufigkeitsklassen von der anderen unterscheidet) in „vermutlich alte Rechtschreibung“ oder „vermutlich neue Rechtschreibung“, die übrigen Korpora in „unklar“. Für die ersten beiden Kategorien wurden je eigene, nach Streuung sortierte Wortformenlisten erstellt. Wortformen, die in einer dieser Listen unter den ersten 100.000 waren, wurden aufgrund ihrer Gesamthäufigkeit in die Zielliste eingemischt, die im letzten Schritt nach dem 100.000sten Eintrag abgeschnitten wurde. Insgesamt wurden 7858 Einträge aufgrund des weicheren Streuungsmaßes hinzugefügt.

2.3 Grundformenlisten

In vielen Fällen sind Wortformenlisten nicht adäquat, z.B. für eine Stichwort(kandidaten)liste im lexikographischen Kontext oder für ähnliche Untersuchungen. In solchen Fällen ist eine Liste von Grund- oder Nennformen geeigneter, wie z.B. DeReWo v-30000g-2007-12-31-0.1 (2007). Die zusätzlichen Problembereiche bei der Erstellung von Grundformenlisten sind in den allgemeinen Anmerkungen zur Reihe DeReWo (DeReWo 2009) ausführlich dokumentiert.

2.4 Relevanz-Sonderfälle

Außer evtl. als Nebeneffekt über die zweistufige Forderung einer ausreichenden Streuung wurde auf Sonderfälle bei dieser Wortformenliste nicht eingegangen.

2.5 Häufigkeitsklassen

Die Häufigkeit einer Wortform 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 Wortformfrequenzen 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 Wortform die Häufigkeitsklasse N, wenn die häufigste Form etwa 2^N-mal häufiger vorkommt als diese Form. Für die Wortformenliste ist der Eintrag mit der höchsten Frequenz *der* mit f(*der*)= 108.861.739, d.h.

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

also $f(wortform) \approx f(der)/2^N$.

Bsp.

N =	0	1	2	3	4	5		10		17
2^N =	2^0	2^1	2^2	2^3	2^4	2^5	…	2^{10}	…	2^{17}
2^N =	1	2	4	8	16	32		1.024		131.072
Bsp.	*der*	*in*	*nicht*	*sind*	*aber*	*Jahr*		*persönlich*		*Tarnkappe*

D.h. *der* ist etwa vier Mal so häufig wie *nicht*, etwa acht Mal so häufig wie *sind* und etwa 131.072 Mal so häufig wie *Tarnkappe*.

In der veröffentlichten Form ist die Liste auch innerhalb der Häufigkeitsklassen nach der

absoluten Häufigkeit sortiert.

2.6 Qualitätskontrolle

Zur Qualitätskontrolle haben wir als integralen Bestandteil des Vorgehens die Randbereiche händisch untersucht. Insbesondere wurden stichprobenartig für bestimmte Wörter die Schreibvarianten vor und nach der Rechtschreibreform systematisch abgeleitet und gegengeprüft.

3 Dateiformat

Die Wortformenliste ist als Datei mit dem Namen DeReWo v-100000t-2009-04-30-0.1 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 Wortformenliste zeilenweise zweispaltig angegeben: Das erste Feld enthält die Wortform, davon mit einem Leerzeichen abgetrennt ist im zweiten Feld deren Häufigkeitsklasse angegeben. In der veröffentlichten Form ist die Liste auch innerhalb der Häufigkeitsklassen nach der absoluten Häufigkeit sortiert.

Referenzen

DeReKo (2009): Deutsches Referenzkorpus, http://www.ids-mannheim.de/kl/projekte/korpora/, Stand: 2009.

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

DeReWo v-30000g-2007-12-31-0.1 (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.

Lizenzbestimmungen

(zu zitieren als:)

> Korpusbasierte Wortformenliste DeReWo, v-100000t-2009-04-30-0.1, mit Benutzerdokumentation,
> http://www.ids-mannheim.de/kl/derewo/,
> © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2009.

Die Wortformenliste, 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.