Benutzerdokumentation

Technical Report IDS-KL-2009-03

zum Produkt

Korpusbasierte Wortgrundformenliste

DeReWo
v-40000g-2009-12-31-0.1

Institut für Deutsche Sprache, Mannheim
Dezember 2009

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.

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-40000g-2009-12-31-0.1

1.1 Was ist DeReWo v-40000g-2009-12-31-0.1?

DeReWo v-40000g-2009-12-31-0.1 ist die Lemmastrecke eines fiktiven Wörterbuchs des öffentlichen Schriftsprachgebrauchs der letzten 30 Jahre im Umfang von 40.000 Lemmata zusammengestellt auf der Grundlage herkömmlicher lexikographischer Kriterien mit besonderer Berücksichtigung der Gebrauchshäufigkeit in Form von Korpusfrequenz.

1.2 Wie wurde DeReWo v-40000g-2009-12-31-0.1 erstellt?

DeReWo v-40000g-2009-12-31-0.1 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

2.1 Korpusbasiertheit

Der Grundformenliste liegen alle Korpora des DeReKo-Archivs Stand Dezember 2009 (DeReKo 2009) zugrunde.

In einem ersten Schritt wurde eine 800.000 Einträge umfassende Wortformenliste aufgrund der Erkenntnisse aus (DeReWo 2009b) erstellt. Die Frequenzangaben in dieser Liste wurden um die Anteile bereinigt, mit denen die Wortformen als Eigennamen verwendet werden. Danach wurde in einem zweiten Schritt eine Grundformenliste abgeleitet, wobei Verfeinerungen der Methodik aus (DeReWo 2007) zum Einsatz kamen.

2.2 Wortformenlisten

Für DeReWo v-40000g-2009-12-31-0.1 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 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 die Wortformenliste wurden die Bestandteile diskontinuierlicher Konstituenten (z.B. Präverbfügungen und Präverben) 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 24 von 273 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*
50:	*dass*	7.574.433	244
60:	*daß*	5.087.207	261

[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

In anderen Fällen reicht die 24/273-Streuung über die Gesamtdatengrundlage nicht aus, um unter den ersten 800.000 berücksichtigt zu werden. Um u.a. diesem Einschnitt des Wortbestandes der deutschen Sprache gerecht zu werden, wurde für die Wörter, 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 800.000 waren, wurden aufgrund ihrer Gesamthäufigkeit in die Zielliste eingemischt, die im letzten Schritt nach dem 800.000sten Eintrag abgeschnitten wurde.

So sinnvoll es ist, über möglichst viele Korpora zu streuen, um Regionalismen, Eigenheiten bestimmter Zeitungen (wie Autorenkürzel oder „proprietäre" Rubrikenüberschriften) zu vermeiden, so wichtig ist es auch, neben den Rechtschreibreformeffekten jungen Wörtern mit hohen Frequenzen eine Chance zu geben, die eine höhere Streuung erst im Laufe der Zeit erreichen werden können. Es bedarf noch eingehenderer Untersuchungen, um herauszufinden, welches Maß bei welcher anvisierten Länge angebracht ist, um nicht den gewünschten Effekt durch das Aufweichen wieder zunichte zu machen. In unserem konkreten Fall haben sich unter den durch diesen zweiten Schritt aufgenommenen Wortformen viele Parallel-Schreibweisen ('ss' in neueren Texten, obwohl 'ß' „richtig" wäre und andersherum) gefunden, aber auch z.B. die Wortform „Schweinegrippe".

2.3 Grundformenlisten

In vielen Fällen sind Wortformenfrequenzen bzw. darauf aufsetzende Wortformenlisten nicht adäquat, z.B. für eine Stichwort(kandidaten)liste im lexikographischen Kontext oder für psychologische Untersuchungen, die auf einem Konzept von „Geläufigkeit" aufbauen. In solchen Fällen ist eine Liste von Grund- oder Nennformen geeigneter. Die Problembereiche bei der Erstellung von Grundformenlisten sind in den allgemeinen Anmerkungen zur Reihe DeReWo (DeReWo 2009) ausführlich dokumentiert.

Die Frequenz einer Nennform, des sogenannten Lemmas, ist die kumulierte Häufigkeit aller Wortformen, die zu dem Paradigma beitragen. Welches jedoch jeweils das Lemma ist und wie es bestimmt wird, lässt sich nur über die Kompetenz eines Sprechers bzw. durch eine tiefe Analyse der jeweiligen Fälle beantworten; beides ist in unseren Szenarien normalerweise nicht möglich, da unser Datenbestand so umfangreich ist, dass diese Aufgaben einerseits nur mit vollautomatischen Verfahren in einem vertretbarem Zeitrahmen bearbeitet werden können. Da wir andererseits an der Authentizität der Daten festhalten und somit keine vermeintlichen Fehler korrigieren, sind vollautomatische Verfahren für eine derartige Analyse utopisch (bestehende Ansätze greifen nicht). Für unsere Zwecke im Rahmen der Reihe DeReWo greifen wir größtenteils

auf die Ergebnisse eines selbst entwickelten automatischen Lemmatisierers (Belica 1994) zurück. In den Bereichen, in denen automatische Verfahren unzuverlässiger sind, haben wir kleine Testmengen bzw. gezielt herausgefilterte Teilmengen von Hand bzw. halbautomatisch ausgewertet. Diese Ergebnisse wurden, soweit möglich, auf die Gesamtheit extrapoliert.

Im Gegensatz zu einem Stamm bzw. zu einer Stammform, die man als Rumpf definiert ohne Flexionsendungen oder sonstige Suffixe oder auch Präfixe, meint man mit Grundform gemeinhin eine Form aus dem Paradigma, die auch realisiert werden kann (was beim Stamm nicht zwingend der Fall ist). Bei Substantiven ist dies standardmäßig der Nominativ Singular, bei Verben der Infinitiv, bei Adjektiven der Positiv.

Von der Angabe einer Grundform wird nur an wenigen Stellen abgewichen, wenn wir der Meinung waren, dass aus der Form nicht ersichtlich ist, welche weiteren Formen sich dahinter verbergen. Als Ersatz geben wir in den Fällen eine entsprechend gekennzeichnete quasi-Stammform an, z.B. *d-* für alle Formen der Artikel *der*, *die*, *das*.

2.3.1 Groß-/Kleinschreibung

Auch wenn bei der Tokenisierung zwischen Groß- und Kleinschreibung unterschieden wurde, haben wir bei der Ableitung der Grundformen diese Entscheidung weitestgehend dem automatischen Lemmatisierer überlassen, der heuristisch aufgrund der Datenlage sich für die eine oder andere Variante entscheidet.

2.3.2 Trennzeichen/Bindestrich

Trenn- bzw. Bindestriche wurden bei der Erstellung der Wortformenliste als Wortformentrenner betrachtet, stehen also für die Ableitung der Grundformen nicht mehr zur Verfügung.

2.3.3 Diskontinuierliche Konstituenten, Präverbfügungen (abtrennbare/abgetrennte Präfixe)

„Der Mensch stammt vom Affen ab. Es stimmt, dass der Mensch vom Affen abstammt. Der Mensch – vom Affen abstammend – ...“

Das Ergebnis des Tokenisierens unterscheidet zwischen den Wortformen *stammt*, *ab*, *abstammt* und *abstammend*, auch wenn im ersten Satz lediglich eine diskontinuierliche Realisierung derselben Grundform „abstammen“ (wie in den anderen beiden Sätzen) zu verzeichnen ist. Uns steht keine Möglichkeit zur Verfügung, diesen Zusammenhang an der Oberfläche exhaustiv zu erkennen.

Bsp.

Lemma	*notwendige „Reparatur“*
abstammen	müsste um das Vorkommen von *stammen* mit abgetrenntem *ab* erhöht werden
ab	müsste u.a. um Vorkommen von *stammen* und abgetrenntem *ab* verringert werden (sowie um evtl. weitere Vorkommen anderer Verben mit dem Präverb *ab*)
stammen	müsste u.a. um Vorkommen von *stammen* und abgetrenntem *ab* verringert werden (sowie um evtl. weitere Vorkommen von *stammen* mit anderem abgetrenntem Präverb)

Als Ersatz bedienen wir uns eines selbst entwickelten Schätzverfahrens („adjVerbFreq"): Eine zusätzliche (im Vergleich zu (DeReWo 2007) deutlich umfangreichere) Informationsquelle gibt uns einen Hinweis darauf, in wie vielen Fällen die vermeintlichen Präverben tatsächlich als echte Präverben zu deuten sind und zu welcher Fügung sie sich dann mit welchem rudimentären Verb verbinden. Die Frequenz der Präpositionen wird dann um die Frequenz der (echten) Präverben reduziert, die Frequenz der Präverbfügungen wird entsprechend erhöht und die Frequenz der rudimentären Verben (quasi die „Nettoverben") analog verringert. Die Vorschläge des Moduls wurden von Hand auf Plausibilität überprüft, insbesondere diejenigen Kandidaten, die sich am stärksten in ihrer Häufigkeitsklasse (s.u.) verändert haben.

Für andere Formen diskontinuierlicher Konstruktionen, wie z.B. *Im-* und *Export*, *auf-* und *abbauen*, müssten wir ähnlich vorgehen, evtl. mit der zusätzlichen (noch zu überprüfenden) Annahme, dass die Häufigkeitsklassen der vervollständigten Formen nicht allzu weit auseinander liegen dürften. Dies wurde für die aktuelle Version allerdings zurückgestellt. Zu beachten ist in der Konsequenz allerdings, dass dadurch die verkürzten Erstglieder derartiger elliptischer Binomiale inklusive des Fugenlauts (falls dieser nicht sonst als Flexionsendung im dem Lemmatisierer bekannten Paradigma des Wortes vorkommt; z.B. „Wirtschafts") in der Grundformliste angeführt sein können.

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

Auch wenn dieser Punkt nicht explizit fokussiert wurde, bewirken zwei Faktoren eine gewisse Bevorzugung eines Quasi-Standards. Die für die Wortformenliste geforderte Streuung sorgt bereits dafür, dass „randständige" Formen eine Hürde zu überwinden haben. Die Heuristiken des Lemmatisierers sorgen dann weiter dafür, dass evtl. „ungewöhnlichere" Formen unter einer „gängigeren" Form als Nennform subsumiert werden. So ist z.B. nicht auszuschließen, dass in manchen Fällen von Rechtschreibneuregelungen mehr als ein Lemma angegeben ist, in anderen Fällen aber alle Formen verschiedener Regelungen unter einem Lemma zusammengefasst sind. Den Zusammenhang zwischen Grundformen, die diskontinuierlich und zeitgleich oder zeitlich versetzt als Variante kontinuierlich verwendet wurden („sitzenbleiben" vs. „sitzen bleiben"), können wir mit unseren Verfahren nicht herstellen.

2.3.5 Neubildungen/Neologismen

Aufbauend auf den Erfahrungen von (DeReWo 2007) haben wir die vom Lemmatisierer nicht erkannten Formen auf partielle Übereinstimmungen mit Adjektiv-, Verb- und Substantivparadigmen überprüft. Die manuelle Auswertung hat z.T. neue Adjektivschreibweisen (z.B. *rau*, *aufwändig*) aufgedeckt, z.T. auch Hinweise auf Adjektive, die – entgegen bisheriger Postulate – (mittlerweile) gesteigert verwendet werden (*politischste*, *universellste*). Die erkannten Neuerungen wurden ggf. als neue Grundform bzw. mit einer zusätzlichen Kumulierung der Häufigkeiten der weiteren Wortformen eingearbeitet.

Bei Verben und Substantiven wurden analog eine Grundform für ein Paradigma mit kumulierten Häufigkeiten eingeführt. Im Vergleich zu Adjektiven war das Verhältnis von Aufwand zu Ertrag in diesen beiden Bereichen deutlich ungünstiger, es konnten deshalb nicht alle Kandidaten betrachtet werden.

Da (echte) Neubildungen aber bisher nicht systematisch erfasst sind, verbirgt sich darin die Gefahr, dass sie zu Unrecht nicht für unsere Auswahl berücksichtigt wurden, obwohl die kumulierte Frequenz ihres Paradigmas die Aufnahme gerechtfertigt hätte.

2.3.6 Adjektivisch gebrauchte Partizipien

Diese Fragestellung wurde für die aktuelle Version den Heuristiken des Lemmatisierers überlassen.

2.3.7 Nennung der Grundform

Diese Fragestellung wurde für die aktuelle Version weitestgehend den Heuristiken des Lemmatisierers überlassen, jedoch wurde halbautomatisch versucht, die vorgeschlagenen Grundformen des Lemmatisierers in ihrer Schreibweise der derzeit gültigen Fassung des ß-Regelung anzupassen.

Für manche Wortformenparadigmen erscheint es wenig sinnvoll, eine Form als Nennform auszuwählen, da hierbei die Wahrnehmung des Unterschieds zwischen Wortform und Grundform unnötig erschwert wird (z.B. *der*, *die*, *das* usw.). Wir haben dafür weitestgehend die Ersatzdarstellung des Lemmatisierers übernommen, die die variierenden Bestandteile durch einen Bindestrich ersetzt (z.B. *d-*, *d-jenig-*). In Einzelfällen wurde diese Ersatzdarstellung von Hand nachgetragen (z.B. *we-* für *wer*, *wem*, *wen*).

2.3.7.1 Movierung

Diese Fragestellung wurde für die aktuelle Version den Heuristiken des Lemmatisierers überlassen, aber nicht explizit hinterfragt.

2.3.7.2 Reflexive Verben

Diese Fragestellung wurde für die aktuelle Version zurückgestellt.

2.3.8 Abgleich mit Wörterbuchlemmastrecken

Eine 60.000 Einträge umfassende Kandidatenliste wurde mit einer Wörterbuchlemmastrecke abgeglichen. Die in dem Wörterbuch nicht belegten Kandidaten wurden speziell im Hinblick auf die u.g. Kriterien der Relevanz-Sonderfälle manuell bewertet und ggf. herausgefiltert.

Desweiteren wurde abschließend ein Abgleich mit der Grundformenliste zu (DeReWo 2007) vorgenommen. Eine erste Überprüfung derjenigen Grundformen, die zwar in jener aber nicht dieser aktuellen Liste vorhanden sind, lässt vermuten, dass es sich fast ausschließlich um derartige Fälle handelt, die aufgrund der Relevanz-Sonderfälle bewusst herausgefiltert wurden bzw. sich in der Häufigkeitsklasse befinden, in der der Einschnitt nach dem letzten Element der Liste gemacht wurde.

2.4 Relevanz-Sonderfälle

2.4.1 Fremdwörter, Anglizismen

Da im Vergleich zu (DeReWo 2007) keine Einschränkung auf ein deutsches Alphabet für die Tokenisierung vorgenommen wurde, konnten Fremdwörter nicht nur der englischen Sprache in die Kandidatenliste gelangen. Ein Großteil der englischen Wörter, die sich nach dem Sprachgefühl des Bearbeiters noch nicht an die deutsche Sprache assimiliert haben, wurde über einen Abgleich mit einer englischen Wortformenliste identifiziert. Darüberhinaus wurde auf englische und andere Fremdwörter ein besonderes Augenmerk bei der Durchsicht nicht in einem Wörterbuch belegter Grundformen gelegt. In beiden Fällen wurde im Zweifelsfall zu Ungunsten des Fremdworts entschieden und darauf verzichtet, die Grundform in der Kandidatenliste zu belassen.

2.4.2 Eigennamen

Bereits auf Wortformebene wurde eine automatische Eigennamenerkennung zur Frequenzkorrektur eingesetzt: Mit Hilfe einer zusätzlichen Informationsquelle wurde ermittelt, wie oft eine Wortform als Eigenname und wie oft als Appellativum verwendet wird. Die in der Wortformenliste verzeichneten Frequenzen wurden darauf jeweils mit dem Anteil des Appellativum-Gebrauchs als Faktor malgenommen. Zusätzlich wurde bei Grundformen, die auf „-er" enden, diese Endung abgespalten. Dieses Zwischenergebnis wurde mit einer Liste von Wortformen abgeglichen, die aufgrund des automatischen Verfahrens zu 100% als Eigenname eingestuft wurden, und von Hand ausgewertet. Den größten Teil dieses Abgleichs machten mit über 400 Wortformen Bezeichner für die Bewohner von Städten oder Regionen aus, die ebenfalls für die veröffentlichte Liste herausgefiltert wurden.

2.4.3 Wortreihen

Diese Fragestellung wurde für die aktuelle Version zurückgestellt.

2.4.4 Kurzwörter

Diese Fragestellung wurde für die aktuelle Version zurückgestellt.

2.4.5 Akronyme, Einzelbuchstaben und Kürzel

Durch einen annähernden regulären Ausdruck wurden Kandidaten für diese Kategorien herausgefiltert. Falls sie nicht auch flektiert in der Wortformenliste belegt sind (Kandidat + Endung -s), wurden sie gelöscht, ansonsten manuell bewertet.

2.4.6 unselbstständige Morpheme

Unselbstständige Morpheme (z.B. *-sche*) wurden – soweit u.a. als Nebeneffekt der Suche nach neuen Paradigmen oder beim Wörterbuchabgleich (s. o.) erkannt – aus der Liste entfernt.

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

Diese Fragestellung wurde für die aktuelle Version zurückgestellt.

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 *d-* mit f(*d-*)= 373.738.420, d.h.

$$N = hk(\textit{grundform}) := \lfloor \log_2(f(\textit{d-})/f(\textit{grundform})) + 0{,}5 \rfloor$$

also $f(\textit{grundform}) \approx f(\textit{d-})/2^N$.

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.	*d-*	-	*und*	*mit*	*als*	*Jahr*		*hören*		*Gulaschsuppe*

D.h. *d-* ist etwa vier Mal so häufig wie *und*, etwa acht Mal so häufig wie *mit* und etwa 131.072 Mal so häufig wie *Gulaschsuppe*.

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.

3 Dateiformat

Die Grundformenliste ist als Datei mit dem Namen DeReWo v-40000g-2009-12-31-0.1 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 Grundformenliste zeilenweise zweispaltig angegeben: Das erste Feld enthält die Grundform, 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

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

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

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 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.

Lizenzbestimmungen

(zu zitieren als:)

> 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.

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.