Allgemeine Anmerkungen

Technical Report IDS-KL-2013-01

zur Reihe

DeReWo

Korpusbasierte Wortlisten

Institut für Deutsche Sprache, Mannheim
Januar 2013

Inhaltsverzeichnis

Vorwort zur Reihe DeReWo

Das Institut für Deutsche Sprache erreichen immer wieder Anfragen nach Häufigkeitsangaben für Wörter, insbesondere werden Listen gewünscht der „*N* häufigsten deutschen Wörter".

Auf Rückfragen, für welche Fragestellung die Angaben genutzt werden sollen, welche Sprachgemeinschaft in welchem Zeitraum gemeint ist, was gezählt und worüber ggf. kumuliert werden soll, welche Einheiten außer acht gelassen werden können oder sollen, ob beim Bestimmen der ersten *N* noch besondere Aspekte zu beachten sind usw. zeigt sich oft erheblicher Klärungsbedarf.

So überflüssig und detailverliebt diese Rückfragen der Sprachwissenschaftler auf eine derart einfache Anfrage auch erscheinen mögen, so folgenreich sind die Antworten für die „korrekte" Beantwortung der ursprünglichen Anfrage. Denn – und das sollte beim Lesen dieser Dokumentation deutlich werden – eine einzige, *die* „Top-N-Liste der deutschen Wörter" gibt es einfach nicht. Die vielen einzelnen, für je eine bestimmte Sprachbetrachtungsperspektive angemessenen Ranglisten unterscheiden sich sowohl in ihrer Zusammensetzung wie auch in der Reihenfolge ihrer Einträge beträchtlich.

Mit der Veröffentlichung der hier dokumentierten Wortlisten unter dem Schlagwort DeReWo bemüht sich das Institut für Deutsche Sprache einen Kompromiss zu finden zwischen der faszinierenden Vielfalt unserer sprachlichen Realität und dem berechtigten Wunsch nach ihrer möglichst kompakten, wenn auch teilweise vereinfachenden Beschreibung. Die DeReWo-Wortlisten sollen nicht den allgemein formulierten Wunsch aus dem ersten Absatz erfüllen, sondern gehen auf ihn – aus einem einzigen der vielen denkbaren Blickwinkel – in wesentlichen Aspekten nur ein. Mithilfe dieser generellen Anmerkungen wollen wir Ihnen einen Überblick über die Problembereiche vermitteln, mit denen wir uns auseinandergesetzt haben; die jeweiligen produktspezfischen Dokumentationen soll Ihnen dabei helfen, die von uns gewählte Betrachtungsperspektive und die daraus resultierenden Folgen für die Interpretation und Handhabung der Liste zu erkennen.

Lesen Sie die Dokumente bitte sorgfältig durch, um unnötige Missverständnisse oder Irrtümer, terminologische Verwechslungen oder methodische Fehlentscheidungen bei der Nutzung der DeReWo -Listen zu vermeiden.

Die DeReWo -Wortlisten sind keine „Freeware", sondern Werke im Sinne des Urheberrechts. Vergewissern Sie sich vor einer Nutzung, dass Sie die Nutzungsbedingungen gelesen und verstanden haben. Eine nicht erlaubte Nutzungsart kann Rechtsfolgen nach sich ziehen.

Download

Die Originale der DeReWo-Wortlisten können unter http://www.ids-mannheim.de/derewo zusammen mit der Dokumentation in der jeweils aktuellen Version abgerufen werden.

1 Grundsätzliches zu DeReWo

1.1 Was ist DeReWo?

DeReWo ist eine Sammlung verschiedener Wortformen- und Grundformenlisten, die unter besonderer Berücksichtigung der Gebrauchshäufigkeit in Form von Korpusfrequenz erstellt werden.

1.2 Wie werden DeReWo-Listen erstellt?

DeReWo-Wortlisten basieren auf dem Deutschen Referenzkorpus DeReKo und werden in einer Kombination von automatischen, semi-automatischen und manuellen Verfahren erstellt, je nachdem, welche Vorgehensweise zur Lösung welcher partiellen Problemstellung sinnvoll bzw. erforderlich ist.

2 Methodik

2.1 Korpusbasiertheit

Die dauerhafte Sicherung einer empirischen Grundlage für die germanistisch-sprachwissenschaftliche Forschung ist eine der zentralen Aufgaben des IDS. Zu diesem Zweck unterhält das Institut seit 1964 eine umfangreiche elektronische Stichprobe deutschsprachiger Texte aus Gegenwart und jüngerer Vergangenheit: das so genannte Deutsche Referenzkorpus (DeReKo 2012).

Das Deutsche Referenzkorpus

- bildet mit über fünf Milliarden Wörtern die weltweit größte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus Gegenwart und neuerer Vergangenheit,
- enthält belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten,
- enthält ausschließlich urheberrechtlich abgesichertes Material,
- ist mit vielen Metainformationen ausgezeichnet,
- wird im Hinblick auf Umfang, Variabilität und Qualität kontinuierlich weiterentwickelt,
- ist zu einem großen Teil kostenlos über die Recherchesoftware COSMAS II zugänglich,
- erlaubt in der Nutzungsphase die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind.

Wir versuchen, mit unseren Korpora den Sprachgebrauch abzubilden. Die Aussagen, die wir aus der Auswertung unserer Korpora ableiten, lassen sich aber aufgrund der Datenzusammensetzung und anderer Faktoren nur bis zu einem gewissen Grad auf „den Sprachgebrauch" verallgemeinern. Allein aber über die Größe des Datenbestandes können wir viele Effekte ausgleichen. Die Frequenz der sprachlichen Einheiten ist für uns ein wichtiges Kriterium für ihre Relevanz.

Dadurch, dass der Datenbestand viele Zeitungstexte enthält, ist er besonders gut geeignet, den „öffentlichen Schriftsprachgebrauch" abzubilden. Bei der Interpretation der Angaben ist jedoch

im Hinterkopf zu behalten, dass der Datenbestand sich über die letzten 30 Jahre erstreckt, dadurch viele Texte in verschiedenen Schreibweisen der verschiedenen Rechtschreibreform-(akzeptanz-)phasen enthält, sowie eine Mischung darstellt von Texten aus der Bundesrepublik Deutschland, aus der ehemaligen DDR, der Schweiz und aus Österreich. Frequenzangaben stellen somit immer eine Kumulation von Frequenzen in Texten mit diesen verschiedenen Eigenschaften dar. Durch den langen Zeitraum wirken z.B. die Häufigkeiten in älteren Texten z.B. bei älteren Realien (wie die Währungsangabe *Mark*) noch nach.

Was unter den zu zählenden sprachlichen Einheiten verstanden werden kann, variiert so sehr, wie der Begriff „Wort" unklar gehandhabt wird.

2.2 Wortformenlisten

Im einfachsten Fall kann man unter „Wort" eine konkrete Wortform verstehen, d.h. die Realisierung einer alphabetischen (oder alphanummerischen) Zeichenkette in einem (hier: geschriebenen) Text. Dazu bedarf es lediglich festzulegen, welche Bestandteile zu einer Wortform gehören, welche Zeichen Wortformen trennen und wie Groß-/Kleinschreibung und evtl. Zweifelsfälle zu handhaben sind.

Für den Fall, dass für die genannten Bereiche Konventionen vereinbart sind, ist es leicht, den Text in die relevanten Zeichenketten zu zerlegen (zu tokenisieren), alle Vorkommen der Zeichenketten (der sogenannten Tokens) durchzugehen und die Häufigkeit aller Zeichenketten mit demselben Erscheinungsbild zu kumulieren. Damit erhält man die Häufigkeit der sogenannten Types. Der Type *zu* hat im vorletzten Satz z.B. die Häufigkeit 3, da drei Tokens dieser Form im Satz gezählt werden.

2.2.1 Groß-/Kleinschreibung

Bereits auf Wortformenebene kann man dafür argumentieren, dass die am Satzanfang groß geschriebene Variante einer ansonsten gleichen, klein geschriebenen Wortform zugeordnet werden sollte. Groß-/Kleinschreibung ist aber doch ein schwächerer Indikator als vermeintlich angenommen: Die Schreibung an Satzanfängen, innerhalb von Titeln, in Zitierungen und diverse stilistische Freiheiten lassen in vielen Fällen alle Kombinationen zu, dass die Variante in ihrer Schreibung eben dieser Wortform oder aber der anderen Variante zugeordnet wird. Analog ist die Fragestellung insbesondere für Binnengroßschreibung und Akronyme zu berücksichtigen.

2.2.2 Trennzeichen/Bindestrich

Je nach Fragestellung kann es sinnvoll sein, den Bindestrich als Wortbestandteil zu betrachten oder auch nicht. Der Trennstrich am Zeilenende wird üblicherweise nicht als Bestandteil einer Wortform gedeutet.

2.2.3 Diskontinuierliche Konstituenten, Präverbfügungen (abtrennbare/abgetrennte Präfixe)

„Der Mensch stammt vom Affen ab. Es stimmt, dass der Mensch vom Affen abstammt."

Oberflächlich ist zwischen den Wortformen *stammt*, *ab* und *abstammt* zu unterscheiden, auch wenn im ersten Satz lediglich eine diskontinuierliche Realisierung derselben Wortes zu verzeichnen ist.

2.2.4 „Ausgewogenheit“

Aufgrund der Zusammensetzung der Datengrundlage kann es sein, dass bestimmte Wortformen im Verhältnis zu dem, was als „Standardsprache“ angenommen wird, über- bzw. unterrepräsentiert sind (u.a. aufgrund Textgenre, Textstil, Regionalismen, abgedeckter Zeit, abgedeckter Themen, Eigennamen).

2.2.5 Rechtschreibreform, insbes. Getrennt-/Zusammenschreibung

Verschiedene Faktoren können die Schreibweisen in den Quellen je nach Stadium und Akzeptanz der Rechtschreibreform beeinflussen.

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.

In einem Flektions- bzw. Wortbildungsparadigma eines „Wortes“ steckt viel Redundanz: Man kann alle wichtigen Informationen zu einer (Nenn-)Form angeben, die Angaben zu allen Wortformen des Paradigmas wären identisch, die Formen unterschieden sich – bis zu einem gewissen Grad – nur durch die unterschiedliche Realisierung aufgrund der Morphologie (o.ä.). Die Frequenz dieser Nennform, des sogenannten Lemmas, ist die kumulierte Häufigkeit aller Wortformen, die zu dem Paradigma beitragen. Für psychologische Untersuchungen wird ein Begriff der Geläufigkeit häufig in Verbindung zur Frequenz gesetzt. Dabei unreflektiert über Wortformen zu argumentieren ist bedenklich: In manchen Paradigmen gibt es z.B. viele homonyme Formen, die dadurch einen Anstieg der Häufigkeit bedingen (s. vor allem auch Formen vom Typ *be-X-t*: *beklagt*, *behauptet*; gleichzeitig 3.PersSingPräsIndAkt, PartPerf und somit auch adjektivisch verwendbar, s.u.); diese werden im Vergleich zu Wörtern mit vielfältigen Formen im Paradigma quasi überbewertet. Unstrittig ist wohl, dass, aber unklar ist noch, inwieweit sich morphologisch (o.ä.) Verwandtes gegenseitig bei „Geläufigkeit“ beeinflusst. Für verschiedene Fragestellungen wäre es deshalb wünschenswert, einen Begriff einer „Grundform“ zu haben, auf den man sich beziehen kann. *Auch wenn wir diesen Begriff nicht vollständig klären bzw. definieren können, wollen wir uns wenigstens in die Richtung bewegen und uns ihm annähern.*

Bei der Definition des Begriffs „Grundform“ wirken zum einen natürlich die Entscheidungen zum Begriff „Wortform“ nach (insbesondere Groß-/Kleinschreibung und Trennzeichen) bzw. müssen hier erneut ausgehandelt werden; zum anderen ergeben sich viele weitere Fragen. Manche dieser Fragen lassen 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 nur mit vollautomatischen Verfahren in einem vertretbarem Zeitrahmen bearbeitet werden können[1]. Da wir an der Authentizität der Daten festhalten und somit keine vermeintlichen Fehler korrigieren, sind vollautomatische Verfahren für eine derartige Analyse utopisch (bestehende Ansätze greifen nicht und wir können aus verschiedenen Gründen dieses Anliegen nicht verfolgen). In manchen Fällen

[1] Ein Verfahren, dass es erforderlich macht, dass jedes Wort in der Datensammlung von einem Menschen betrachtet und bewertet wird, würde, selbst wenn nur eine Sekunde pro Wort aufgewendet werden müsste, bei dem heutigen Umfang bereits 111 Jahre in Anspruch nehmen.

versuchen wir, kleine Testmengen von Hand bzw. halbautomatisch auszuwerten. Diese Ergebnisse werden, soweit es geht, auf die Gesamtheit extrapoliert.

Im Gegensatz zu einem Stamm bzw. zu einer Stammform, die man als Rumpf definiert ohne Flektionsendungen 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. Bereits dabei ergeben sich Probleme: Sind adjektivisch gebrauchte Partizipien auf den Positiv oder auf den Infinitiv des zugrunde liegenden Verbs zurückzuführen? Präverbfügungen (früher auch „präfigierte Verben") manifestieren sich in Texten z.T. als kontinuierliche, z.T. aber auch als diskontinuierliche Elemente. Letztere sollten sicher derselben Grundform (dem Infinitiv der Präverbfügung) zugeordnet werden wie die ersteren, stellen aber ein großes Problem für die automatische Erkennung dar.

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, kann es sinnvoll sein, für die Bestimmung der Grundform davon zu abstrahieren. Für die Bestimmung einer Grundform ist die Unterscheidung ein wesentlich unzuverlässigerer Indikator als gemeinhin angenommen werden könnte: Am Satzanfang, in Überschriften, bei Titeln, festen Wendungen (*Tag der Offenen Tür*, vgl. Rechtschreibreform), stilistischen Spielereien, bei Zitaten und Hervorhebungen wird durchaus von der üblichen Variante abgewichen. Dabei wird nicht nur groß- statt kleingeschrieben. Das Internet macht sich auch hier bemerkbar: Viele Varianten fließen auch durch die Bezugnahme auf WWW- und Email-Quellen und die Angabe der entsprechenden Adressen (die viele Wörter der deutschen Sprache in Kleinschreibung enthalten) in die Sprache ein. Mit dem gleichen Effekt, aber ebenfalls für uns nicht ohne weiteres oberflächlich erkennbar, werden Konversionen in der Sprache gebraucht (Nullderivationen: substantivierte Verben und Adjektive, Denominalisierungen), bei denen evtl. darüber nachgedacht werden müsste, ob für diese eigene Grundformen anzusetzen sind.

Mittel- und langfristig wäre zu überlegen, die quantitativen Verteilungen innerhalb der vermeintlichen Paradigmen mit zu der Entscheidung heranzuziehen, wie viele bzw. welche der Formen als Grundformen anzusetzen sind. Interessant erscheint auch die Idee eines kontrastiven Vergleichs der Kookkurrenzprofile der verschiedenen Wortformen eines Lemmas.

2.3.2 Trennzeichen/Bindestrich

Auch wenn Trenn- bzw. Bindestriche als Bestandteile einer Wortform gedeutet werden, ist zu entscheiden, ob die verschiedenen Formen mit und ohne Striche nicht als zu einem Paradigma und somit zu einer Grundform gehörig betrachtet werden sollten.

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 Präfix *ab*)
stammen	müsste u.a. um Vorkommen von *stammen* und abgetrenntem *ab* verringert werden (sowie um evtl. weitere Vorkommen von *stammen* mit anderem abgetrenntem Präfix)

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.

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

Je nachdem, wie weit der Grundformenbegriff ausgelegt werden soll, könnten auch Schreibvarianten unter demselben Lemma subsumiert werden: *daß/dass*, *aufwendig/aufwändig*, usw. Damit öffnet sich ein weites Feld regionaler (Schweizer *dass*) und diachroner Varianten. Auch internet-typische Schreibweisen und die verschiedenen (Akzeptanz-)Phasen der Rechtschreibreform haben einiges zu einer Dynamik bzw. Uneinheitlichkeit des Sprachgebrauchs beigetragen. Gerade auch der Bereich der Getrennt-/Zusammenschreibung ist eng verflochten mit der Frage, welche Einheit als Lemma anzusetzen ist (*sitzenbleiben* vs. *sitzen bleiben* vs. *er blieb sitzen* usw.).

Mit dem letzten Punkt berühren wieder die Fragen, die wir bereits bei den diskontinuierlichen Konstituenten nur ansatzweise aufrollen konnten. Quantitative Vergleiche zwischen den hier genannten Schreibweisen sind deshalb aber nur schwerlich möglich, da die spezifische Information bei der Kumulierung verlorengegangen sein kann.

2.3.5 Neubildungen/Neologismen

Unser Bestreben ist es, mit unseren Korpora ein Abbild des Sprachgebrauchs und natürlich auch von dessen Wandel nachzuzeichnen. Der Begriff einer „Grundform" stößt dabei an gewisse Grenzen, weil wir damit vorwegnehmen müssen, für welche verschiedenen Formen eine vom Sprecher/Schreiber intendierte gemeinsame Grundform anzunehmen ist. Manche Paradigmen von neuen Wortformen lassen sich leicht systematisch erschließen. Durch die Rechtschreibreform sind neue Schreibweisen eingeführt worden, die Flektion entspricht aber weitestgehend den alten Regularitäten. In den Fällen kann man versuchsweise die automatisch nicht erkannten Formen auf partielle Übereinstimmungen mit Flektionsparadigmen überprüfen. Die manuelle Auswertung kann z.T. neue Adjektivschreibweisen (z.B. *rau*, *aufwändig*) aufdecken, z.T. auch Hinweise auf

Adjektive, die – entgegen bisheriger Postulate – (mittlerweile) gesteigert verwendet werden (*politischste*, *universellste*). Die erkannten Neuerungen können ggf. als neue Grundform bzw. mit einer zusätzlichen Kumulierung der Häufigkeiten der weiteren Wortformen eingearbeitet werden.

Bei Verben kann entsprechend eine Grundform für ein Verbparadigma, bei Sustantiven eine Grundform für ein Substantivparadigma mit kumulierten Häufigkeiten eingeführt werden.

Da Neubildungen bisher nicht systematisch erfasst sind, verbirgt sich darin die Gefahr, dass sie zu Unrecht nicht für unsere Auswahl berücksichtigt werden, obwohl die kumulierte Frequenz ihres Paradigmas die Aufnahme rechtfertigt.

2.3.6 Adjektivisch gebrauchte Partizipien

Bei vielen adjektivisch gebrauchten Partizipien ist noch vollkommen transparent, von welchem Verb sie abgeleitet wurden, etwa *das entspannende Bad*. Obwohl flektiert eindeutig als Adjektiv erkennbar, spüren wir beim Lesen noch die große Nähe zu dem Verb *entspannen*. Anders bei gewissermaßen verblassten Bildungen wie *das spannende Buch*. Es ist durchaus nachvollziehbar, wenn wir intuitiv einmal das Adjektiv *spannend* (genaugenommen das Partizip Präsens des Verbs *spannen*), das andere Mal aber das Verb *entspannen* als Grundform ansetzen würden. In anderen Fällen ist das Partizip (Perfekt) homonym zu anderen Formen des Paradigmas, teilweise zur dritten Person Singular (*er (hat) behauptet*) oder zur dritten Person Plural/Infinitiv (*wir verlaufen uns/es ist nicht schwer, sich zu verlaufen/wir haben uns verlaufen*), so dass unklar ist, wenn wir diese als Grundformen ansetzen, wie welche Wortformen auf diese verteilt werden können.

Auch hier wäre es lohnenswert, sich das Kookkurrenzverhalten des Wortformen anzuschauen, da sich die Heterogenität (als Rechtfertigung für die Ansetzung mehrerer Grundformen) in einer Partitionierung des Kontextverhaltens zeigen sollte.

2.3.7 Nennung der Grundform

In den meisten Fällen ist es, wie in der Einleitung bereits kurz erwähnt, plausibel, eine konkrete unmarkierte Form aus dem Paradigma als Grundform anzugeben. Für Nomen ist dies der Nominativ Singular. Einige Spezialfälle, die nur im Singular verwendet werden. stellen keine Problemfälle dar (sog. Singularia Tantum wie z.B. *Hass*, *Milch*, *Sand*, *Getreide*, *Obst*), im komplementären Fall der Nomen, die nur im Plural vorkommen (sog. Pluralia Tantum wie z.B. *Eltern*, *Geschwister*, *Spesen*, *Ferien*) erscheint es aber intuitiv sinnvoll, von dieser Regelung abzuweichen. Die Frage wird allerdings komplexer, weil analog hinterfragt werden müsste, ob nicht bei allen Paradigmen eine fast ausschließlich (oder sogar generell: die am häufigsten) verwendete Form angegeben werden sollte.

Eine ähnliche Frage stellt sich auch bei Adjektiven, die gemeinhin im Positiv genannt werden. Diejenigen, die nicht steigerungsfähig sind, stellen wiederum kein Problem dar (*einzig*, *ewig*, *erster*, *letzter* (*letzterer*??)). Diejenigen allerdings, die nur gesteigert vorkommen (*weltgrößte*, *dienstältere*, *dienstälteste*, *äußerst* (von „*außen*“??)), sollten mittelfristig entsprechend gekennzeichnet werden.

Für manche Wortformenparadigmen halten wir es aber für nahezu kontraintuitiv, 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 werden dafür weitestgehend

eine Ersatzdarstellung übernehmen, die die variierenden Bestandteile durch einen Bindestrich ersetzt (z.B. *d-*, *d-jenig-*). In Einzelfällen kann diese Ersatzdarstellung von Hand nachgetragen sein.

2.3.7.1 Movierung

Für Personen-, insbesondere für Berufsbezeichnungen gibt es häufig zwei parallele Formen (männlich/weiblich). Einige Formen lassen sich oberflächlich eindeutig jeweils einem Geschlecht zuordnen (z.B. *Lehrer/Lehrerin*, *Witwe/Witwer*), einzelne Formen sind aber homonym sozusagen auf beiden Seiten vertreten (*Beamte/Beamter* (*der/die/eine* vs. *ein*)). Unklar ist, zu welcher Nennform die verschiedenen Formen subsumiert werden sollen.

2.3.7.2 Reflexive Verben

Lexikographisch ist es üblich, bei reflexiven Verben das Reflexivpronomen in der dritten Person mit zum Infinitiv bei der Nennform mit anzugeben. Aus verschiedenen Gründen sehen wir normalerweise davon ab (Schwierigkeit der Erkennung, vgl. diskontinuierliche Konstituenten) und zählen die Pronomen und Verben getrennt. Bei der Interpretation der jeweiligen Frequenzangaben sollte dies jeweils bedacht werden.

2.3.7.3 Pronomen

Gerade bei Pronomen gibt es ein Vielfalt von Entscheidungsmöglichkeiten, entlang welcher Dimensionen Wortformen zu Lemmata zusammengefasst werden. Dass die flektierten Formen attributiv gebrauchter besitzanzeigender Fürwörter zu einem zusammengefasst werden, liegt noch nahe (mein, meine, meines, meinen, meinem, meiner zu mein-). Offen ist aber bereits, ob prädikativ gebrauchte (meins) mit dazugehören. Genauso unklar ist, ob in der Dimension Person (mein vs. dein), zwischen den verschiedenen Genus (sein vs. ihr) und/oder (etwa bei Personalpronomen) Kasus (ihn vs. ihm) unterschieden werden soll.

Es könnte ein Lemma für alle Personalpronomen angesetzt werden. Üblich sind aber eher sechs verschiedene, je nach Person/Numerus-Kombination, oder acht, je nach Person/Numerus/ Genus-Kombination. Je nach Überzeugung gibt es viele unterschiedliche Konventionen, auch gerade bei der Angabe der Nennform.

2.3.8 Abgleich mit Wörterbuchlemmastrecken

Grundsätzlich erscheint es natürlich sehr vielversprechend, eine Grundformenliste mit anderen Lemmalisten abzugleichen. Allerdings gibt es in diesen Listen Lücken und „Leichen“, und der Bezugszeitpunkt kann anders sein: Neueste Formen wurden evtl. nicht berücksichtigt (Neologismen), z.T. liegen andere Prinzipien zugrunde: Fremdwörter, Fachbegriffe, veraltete Schreibweisen, unselbstständige Morpheme, Einzelbuchstaben und Akronyme, verschiedene Epochen gerade während der verschiedenen Phasen der Rechtschreibreform werden je nach Vorgaben unterschiedlich gehandhabt.

	in unserer Liste	*nicht in unserer Liste*
im Wörterbuch	leichte Bestätigung, könnte im Wörterbuch aber erst viel später berücksichtigt sein	könnte ebenfalls im Wörterbuch weiter hinten zu finden, oder, sofern uns kein Fehler unterlaufen ist, ein veralteter Eintrag (eine „Leiche") sein
nicht im Wörterbuch	sofern uns kein Fehler unterlaufen ist, könnte dies ein fehlender Eintrag (eine Lücke) sein (*)	leichter Hinweis auf geringe Relevanz, könnte aber auch von beiden zu unrecht als zu niedrig relevant eingestuft worden sein

Leider unterscheiden sich die verschiedenen Listen normalerweise auch in ihrem Umfang und es liegt kein explizites Relevanzkriterium vor (wie bei uns primär die Häufigkeit), so dass wir die Vergleichslisten nicht sortieren können. Wir wissen also nicht, welche Einträge die ersten *N* wären.

Die für unsere Listen relevantesten Erkenntnisse erhoffen wir uns aus der Überprüfung derjenigen Formen, die zwar in unseren Listen, aber nicht im Wörterbuch verzeichnet sind (*). Diese können wir manuell in zwei Mengen aufteilen: Diejenigen, die aus unserer Sicht einen relevanten Eintrag darstellen, und diejenigen, die als nicht-relevant herauszufiltern sind. Letztere Menge bedarf jeweils einer detaillierteren Auswertung.

2.4 Relevanz-Sonderfälle (verstärkend oder abschwächend, „Häufigkeit" relativierend)

2.4.1 Fremdwörter, Anglizismen

Auch wenn Fremdwörter, insbesondere Anglizismen sehr häufig in der deutschen Sprache gebraucht werden, ist doch nicht unumstritten, inwieweit sie zu einem Kernbestand der deutschen Sprache gehören.

Entscheidungen bei der Erstellung der Wortformenliste haben evtl. Einfluss darauf, ob authentische Schreibweisen von Fremdwörtern berücksichtigt werden oder nicht. Falls die Wortformenliste z.B. nur das deutsche Alphabet zugrunde legt, die Fremdsprachen aber abweichende Alphabete benutzen (z.B. das Französische mit diakritischen *e*: *é*, *è*, *ê*), werden die authentischen Schreibweisen übergangen, im Gegensatz zu Fremdwörtern aus Sprachen, die dasselbe Alphabet benutzen (wie z.B. das Englische)

2.4.2 Eigennamen

Wenn wir annehmen, dass wir die ersten 10.000, 20.000 oder 30.000 Einträge einer Wortliste bestimmen können, gehen wir auch davon aus, dass wir die Listen nach Gewichtigkeit sortieren können. Unser wichtigstes Kriterium dafür ist die Frequenz. Unter den ersten *N* Einträgen finden sich allerdings viele Formen, die man im weitesten Sinne als Eigennamen auffassen kann. Auch unabhängig von der Zusammensetzung unserer Datengrundlage (viele Zeitungen, in denen viel über Personen aus Politik, Wirtschaft, Kultur und Sport berichtet wird), erscheint fragwürdig, ob unsere Auswahl auch mit einer gedämpften Anzahl an Eigennamen den Bestand dessen wiedergibt, was wir als Kern einer Sprache betrachten würden. Um eine Sprache zu erlernen,

braucht man dieses große Repertoire an Eigennamen sicher nicht; um eine Zeitung lesen zu können evtl. schon, nur ist in dem Kontext auch selbstverständlich, dass dazu andere Wissensquellen herangezogen werden müssen. Unsere Idee ist deshalb, ggf. separate Listen der verzeichneten Eigennamen ergänzend zu den Wortlisten herauszugeben.

Die Erkennung der Eigennamen ist verhältnismäßig vielleicht sogar das „kleinere Problem". Ähnlich wie bei den diskontinuierlichen Konstituenten können wir uns einer weiteren Informationsquelle bedienen. Alternativ können wir mit einem Verfahren uns nach und nach zu einer präziseren Eigennamenmenge hochschaukeln: Beginnend mit Wortformen, die relativ eindeutig meistens vor Eigennamen stehen (*Herr*, *Frau*, *Prof.*, *Präsident*, *Kanzler*, *Vorsitzende* usw.) lassen sich die danach folgenden Wortformen zunächst quantitativ und dann qualitativ auswerten. Mit ihrer Hilfe kann man in die andere Richtung zurück schauen, welche Wortformen typischerweise davor stehen. Dieser Prozess lässt sich ggf. mehrmals wiederholen, bis sich die Menge stabilisiert. Ein zweites, und sehr wahrscheinlich das größere Problem ist allerdings die Bewertung, wie oft tatsächlich eine Wortform als Eigenname verwendet wird (*Kohl* vs. *Kohl*) und auf wie viele verschiedene Instanzen damit referiert werden kann. Denn erst daraus lässt sich ein Zusammenhang zwischen Frequenz und Relevanz ableiten. Um diese Frage zu beantworten, könnte man (a) kleine Testmengen von Hand auswerten und dann versuchen, zu extrapolieren oder (b) sich das Kookkurrenzverhalten der Wortformen anschauen, da sich die Heterogenität (Referenz auf viele Personen bzw. auf Person und Objekt) in einer Partitionierung des Kontextverhaltens zeigen sollte.

2.4.3 Wortreihen

Manche Wörter (egal welcher Wortklasse) gehören zu kleinen überschaubaren, in sich geschlossenen Reihen: Wochentag- und Monatsbezeichnungen, Tageszeiten, Farben usw. Einerseits ist nicht grundsätzlich anzunehmen, dass alle Elemente einer Reihe ungefähr gleich häufig vorkommen. Es ist z.B. durchaus plausibel, über Primärfarben und im Straßenverkehr und der Politik verwendete Ausprägungen (*rot*, *gelb*, *blau*, *grün*) öfter zu sprechen als über andere; andererseits kann es durch Eigenschaften der Datengrundlage und die Willkür des Einschnitts (der zumal mitten in einer Häufigkeitsklasse sein kann) passieren, dass ein Teil der Reihe in unserer Auswahl ist, ein anderer Teil jedoch nicht. Hier ist es überlegenswert, zumindest alle Elemente der Reihe der noch berücksichtigten Häufigkeitsklasse (evtl. sogar einer darüber hinaus) bevorzugt zu berücksichtigen, oder ihre Grenzwertigkeit in einer zusätzlichen Liste festzuhalten.

2.4.4 Kurzwörter

Zu manchen Wortformen gibt es verkürzte Varianten (z.B. *Alu* für *Aluminium*), bei denen sich auch die Frage stellt, ob sie gemeinsam unter eine Grundform gefasst werden sollen. Gerade darin zeigt sich die Veränderung des Sprachgebrauchs, heute weiß kaum noch ein Sprecher, dass *Bus* nur die Kurzform von *Omnibus* ist. Dieses Phänomen müsste aber sehr wahrscheinlich aus diachroner Perspektive betrachtet werden, interessant wäre hierbei sicher auch, einen Zusammenhang zwischen den Längen der jeweiligen Formen und ihren Gebrauchshäufigkeiten herzustellen.

2.4.5 Akronyme, Einzelbuchstaben und Kürzel

Uns erscheint es fraglich, ob Akronyme – ähnlich wie Eigennamen – als zum Kernbestand einer Sprache gehörig zu betrachten sind, oder ob sie stattdessen evtl. in eigenen Listen dokumentiert werden sollten. Bei Einzelbuchstaben/Kürzeln stellt sich die Frage, ob sie eigene Einheiten darstellen oder zu ihrer Vollform expandiert werden sollen (z.B. *z.B.* zu *zum* und *Beispiel*).

2.4.6 Unselbstständige Morpheme

Manche Wörterbücher verzeichnen gezielt auch unselbstständige Morpheme (z.B. *-sche*). Ein Vorkommen in unseren Listen ist möglich, dann aber eher zufällig, da sie nur sehr schwierig systematisch zu erfassen sind, sei es, um sie alle explizit zu berücksichtigen oder herauszufiltern.

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

In Fällen der Verschmelzung stellt sich die Frage, ob die beteiligten Formen (*ans* = *an das*, *zum* = *zu dem*, ...) nicht herausgebrochen werden und entsprechend behandelt werden sollten.

2.5 Häufigkeitsklassen

Die Häufigkeit einer Wort- oder 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 Wortformfrequenzen ergeben, die oben beschriebenen Problemfelder können zusätzliche Verschiebungen bei den Grundformfrequenzen bewirken. Als relativ stabil und aussagekräftig – gerade auch beim Vergleich unterschiedlich großer Datenbestände – hat sich erwiesen, Häufigkeiten in Form von Häufigkeitsklassen anzugeben[2]. Dabei hat ein Wort die Häufigkeitsklasse N, wenn das häufigste Wort etwa 2^N-mal häufiger vorkommt als dieses Wort.

$$N = hk(wort) := \lfloor \log_2(f(häufigstes\_wort)/f(wort)) + 0{,}5 \rfloor$$

also $f(wort) \approx f(häufigstes\_wort)/2^N$.

Sofern nichts anderes angegeben ist, sind in den Listen im Originalzustand die Einheiten auch innerhalb der Häufigkeitsklassen nach der absoluten Häufigkeit sortiert.

2.6 Qualitätskontrolle

Zur Qualitätskontrolle werden als integraler Bestandteil des Vorgehens soweit möglich Randbereiche und besondere Phänomene händisch untersucht.

3 Ausblick

Auf der Grundlage der Erfahrungen, die wir bei jeder neuen Version sammeln, und aufgrund der Rückmeldungen, die uns erreichen, planen wir Neuauflagen unter jeweils festzulegenden Gesichtspunkten und in jeweils festzulegenden Größenordnungen.

Referenz

DeReKo (2012): Deutsches Referenzkorpus, http://www.ids-mannheim.de/kl/projekte/korpora/, Stand: 2012.

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

Lizenzbestimmungen

(die jeweiligen Wortlisten sind instantiiert zu zitieren als:)

> Korpusbasierte Wortformenliste (bzw. Wortgrundformenliste) DeReWo, <v-xxx>, mit Benutzerdokumentation,
> http://www.ids-mannheim.de/derewo,
> © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, <Jahresangabe>.

Die Wortlisten, ihre Dokumentation und diese generellen Anmerkungen bilden eine Einheit. Diese Lizenzbestimmung darf aus keinem der Dokumente entfernt werden.

Die Werke sind jeweils unter die Creative Commons-Lizenz (by-nc) gestellt (http://creativecommons.org/licenses/by-nc/3.0/deed.de).

Namensnennung – Keine kommerzielle Nutzung 3.0 Unported

Sie dürfen:

- das Werk vervielfältigen, verbreiten und öffentlich zugänglich machen
- Bearbeitungen des Werkes anfertigen

zu den folgenden Bedingungen:

- Namensnennung. Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen (wodurch aber nicht der Eindruck entstehen darf, Sie oder die Nutzung des Werkes durch Sie würden entlohnt).
- **Keine kommerzielle Nutzung.** Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden.
- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter welche dieses Werk fällt, mitteilen. Am Einfachsten ist es, einen Link auf die o.g. Seite einzubinden.
- Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die Einwilligung des Rechteinhabers dazu erhalten.

Diese Lizenz lässt die Urheberpersönlichkeitsrechte unberührt.

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.