universität wien

Diplomarbeit

Vergleich von Worthäufigkeitsangaben in CELEX und DeReKo

Verfasserin

Karin Gmoser

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, August 2013

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuer: Ao. Univ.-Prof. Mag. Dr. Alfred Schabmann

INHALT

1 EINLEITUNG

In der einschlägigen Literatur zum Thema Worterkennung ist der Worthäufigkeitseffekt einer der stabilsten und am häufigsten thematisierten Effekte (Cortese & Balota, 2012). Versuchspersonen reagieren bei häufig auftretenden Wörtern schneller und/oder exakter als bei seltenen Wörtern – dies betrifft Aufgaben wie etwa lexikalische Entscheidung (Balota & Chumbley, 1984), Wortbenennung (Grainger, 1990) und tachistoskopische Identifizierung (Howes & Solomon, 1951).

Um jedoch den Worthäufigkeitseffekt zu untersuchen bzw. zu kontrollieren, sind zuverlässige Angaben darüber, wie oft ein Wort tatsächlich in der Sprache vorkommt, unerlässlich (Brysbaert & New, 2009).

Zevin und Seidenberg (2002) verglichen für den englischsprachigen Raum Worthäufigkeitsangaben aus unterschiedlichen Datenbanken anhand dreier großer Studien zur Worterkennung. Schlussfolgernd weisen sie darauf hin, dass schlechte Häufigkeitsangaben zur Verzerrung von Untersuchungsergebnissen führen können, wenn die interessierende Variable mit der Variable Worthäufigkeit korreliert ist. In diesen Fällen ist der vermeintlich gefundene Effekt lediglich ein maskierter Worthäufigkeitseffekt (Zevin & Seidenberg, 2002).

Burgess und Livesay (1998) verglichen das nur knapp über eine Million Wörter umfassende Brown Corpus (Kucera & Francis, 1967) mit dem bedeutend größeren – 131 Millionen Wörter - und aktuelleren HAL-Korpus. Bei hochfrequenten Wörtern zeigte sich eine relativ hohe Übereinstimmung (r=.96) der beiden Korpora, bei niederfrequenten Wörtern jedoch lag die Übereinstimmung nur noch zwischen r=.12 und r=.14 (Burgess & Livesay, 1998).

Brysbaert und New (2009) konnten nachweisen, dass ein Korpusumfang von 1 bis 3 Millionen Wörtern ausreicht, um reliable Häufigkeiten für hochfrequente Wörter zu

ermitteln, die Häufigkeiten niederfrequenter Wörter (Häufigkeit kleiner als 10/Million) erfordern jedoch einen Korpusumfang von mindestens 16 Millionen Wörtern, um zuverlässige Schätzungen zu erlangen.

In ihrer Arbeit gehen Brysbaert und New (2009) nicht nur auf den Umfang, sondern auch auf die Qualität der verwendeten Korpora ein. Sie betonen die Wichtigkeit der verwendeten Quellen und kritisieren die einseitige Zusammensetzung vieler Korpora. Einer der Kritikpunkte ist, dass in den meisten Fällen Bücher, Zeitungen und Magazine die Hauptquellen darstellen. Brysbaert und New (2009) weisen auf einige Probleme hin, die diese Zusammenstellung mit sich bringt:

- Texte dieser Art werden üblicherweise editiert und so sprachlich verbessert.
- Da Wortwiederholungen vermieden werden, kommt es zu einer übertriebenen lexikalischen Variation.
- Es werden oft Themen behandelt, die nicht das alltägliche Leben wiederspiegeln.

D.h. aus Büchern und Zeitschriften erstellte Korpora spiegeln nicht die alltagssprachliche Realität! Sowohl Burgess und Livesay (1998) als auch Brysbaert und New (2009) empfehlen als Alternative die Erstellung von Korpora aus Quellen wie etwa Internet-Newsgroups, Online-Diskussionsforen oder Untertiteln von Filmen und TV-Beiträgen, da diese eher dem alltäglichen Spracherleben des durchschnittlichen Menschen gerecht werden.

Blair, Urland und Ma (2002) generierten Worthäufigkeitsschätzungen aus Internetsuchmaschinen und verglichen diese unter anderem mit den Häufigkeitsangaben aus dem englischsprachigen Korpus von CELEX. Diese zeigten generell eine relativ gute Übereinstimmung, jedoch nicht bei Wörtern, die verhältnismäßig neu im Wortschatz sind oder öfter in gesprochener als geschriebener Sprache vorkommen (Blair, Urland, & Ma, 2002).

Die vorliegende Diplomarbeit soll klären, ob sich die Worthäufigkeitsangaben in der Korpusdatenbank des Centre for lexical Information (CELEX) signifikant von den Häufigkeitsangaben im DeReKo – dem Deutschen Referenzkorpus des Instituts für Deutsche Sprache Mannheim (IDS) – unterscheiden. Die Ergebnisse von Brysbaert und New (2009) führen zu der Hypothese, dass die Häufigkeitsangaben in CELEX aufgrund des geringen Umfangs und der relativ unausgewogenen Quellen des Korpus von geringerer Qualität sind als die Angaben aus dem wesentlich größeren und aktuelleren Korpus des DeReKo. Es soll ferner auch überprüft werden, ob eine mögliche Abweichung der Häufigkeitsangaben systematisch bei hoch- oder niederfrequenten Wörtern auftritt, und ob CELEX im Falle einer Abweichung eher zu einer Überschätzung oder einer Unterschätzung der Häufigkeitsangaben tendiert.

1.1 METHODIK

Aus CELEX wurde eine Zufallsstichprobe von 285 Wörtern – unterteilt in vier Gruppen nach logarithmierter Häufigkeit pro Million - gezogen. In weiterer Folge wurde mithilfe der vom IDS Mannheim zur Verfügung gestellten Software COSMASII zu jedem Wort die dazugehörige Häufigkeit im DeReKo ermittelt. In SPSS wurde für jedes Wort die Log-Likelihood-Ratio berechnet, welche angibt, wie stark die Häufigkeitsangaben voneinander abweichen.

1.2 AUFBAU DER ARBEIT

Einleitend werden die wichtigsten Begriffe der quantitativen Linguistik erklärt. Im Anschluss wird kurz auf die Geschichte der Korpuslinguistik eingegangen. Es wird beschrieben, wie es zu den ersten Häufigkeitszählungen im deutschsprachigen Raum kam und was hier der neueste Stand der Forschung ist. Es werden einige Studien aus dem englischsprachigen Raum präsentiert, in welchen nachgewiesen werden konnte,

dass sich der geringere Umfang und veraltete Quellen von Korpusdatenbanken negativ auf die Qualität von Worthäufigkeitsangaben auswirken. Weiters erfolgt eine Erklärung der Besonderheiten bei der Verteilung von Worthäufigkeiten: diese sind nie normalverteilt, sondern folgen einer sogenannten Zipf'schen Verteilung, d.h. es gibt einige wenige hochfrequente Wörter und eine enorm große Anzahl an niederfrequenten Wörtern.

Im nächsten Kapitel wird auf den Worthäufigkeitseffekt eingegangen. Es erfolgt ein kurzer Überblick über Modelle zur Wahrnehmung häufiger Wörter. Danach werden einige ausgewählte Studien präsentiert, die sich mit dem Worthäufigkeitseffekt auseinandersetzen. Weiters soll es einen Überblick über den aktuellen Stand der Forschung geben. Ziel des Kapitels ist es, zu verdeutlichen, dass es für viele Untersuchungsdesigns unerlässlich ist, den Worthäufigkeitseffekt als mögliche Stör-/Moderatorvariable zu kontrollieren.

In einem weiteren Kapitel wird der grundsätzliche Aufbau von Korpusdatenbanken erklärt, und einige große Korpusprojekte des deutschen Sprachraums vorgestellt.

In den darauf folgenden Kapiteln wird näher auf die untersuchten Datenbanken eingegangen: die Entstehungsgeschichte sowie der genaue Aufbau und die Konstruktionsweise von CELEX und DeReKo werden – soweit bekannt - erläutert, ebenso die zugrundeliegenden Quellen. Hier wird auch beschrieben, welche Schwierigkeiten und Fehlerquellen bei der Korpuserstellung auftreten, beispielsweise der Umgang mit homographischen Wörtern (= Wörter mit gleicher Schreibweise aber unterschiedlicher Bedeutung).

In den Kapiteln des empirischen Teils werden zunächst das Untersuchungsziel und die dazugehörigen Hypothesen erläutert. Anschließend erfolgt eine genaue Beschreibung des Untersuchungsdesigns. Weiters wird die Datenerhebung und Zusammenstellung der Stichprobe erläutert.

Im nächsten Abschnitt wird anhand der Arbeiten von Rayson und Garside (2000) und Dunning (1993) argumentiert, dass für den Vergleich von Worthäufigkeiten die Log-Likelihood-Ratio anderen Verfahren – wie etwa Pearsons Chi-Quadrat – vorzuziehen ist.

Danach folgt die genaue Erläuterung der vorgenommenen Berechnung der Log-Likelihood-Werte, und im Anschluss daran erfolgt die Darstellung der Ergebnisse der Log-Likelihood-Analyse.

In der abschließenden Diskussion werden die Ergebnisse interpretiert, die Erhebung und Berechnung kritisch beleuchtet und ein Ausblick in die Zukunft bzw. eine Empfehlung zur Verwendung der gewonnenen Erkenntnisse gegeben.

2 THEORETISCHER TEIL

Das folgende Kapitel soll grundlegende Begriffe der quantitativen Linguistik erklären und einen Überblick über bisherige Forschungsarbeiten in diesem Bereich geben.

2.1 QUANTITATIVE LINGUISTIK

Quantitative Linguistik: „beschäftigt sich unter Verwendung statistischer Methoden mit der kontrollierten Untersuchung sprachlicher Regularitäten unter quantitativen Aspekten. Dient u.a. der Herstellung von Häufigkeitswörterbüchern“ (Bußmann, 2008, S.565)

Häufigkeitswörterbuch: „auch Frequenzwörterbuch. Statistische Registrierung der häufigsten Wörter einer Sprache, die aufgrund quantitativer Kriterien als Wörter mit dem höchsten Gebrauchswert ausgewählt werden. Solche lexikographischen Frequenzuntersuchungen basieren auf einer als repräsentativ angesehenen breiten Streuung verschiedener Textsorten.“ (Bußmann, 2008, S.253)

Korpus: „Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten liegen üblicherweise in digitalisierter Form vor. Bestandteile des Korpus sind die Daten selbst, und möglicherweise auch Metadaten, welche diese Daten beschreiben, und linguistische Annotationen welche diesen Daten zugeordnet sind.“ (Lemnitzer & Zinsmeister, 2010, S.8)

„Stellenwert und Beschaffenheit des Korpus hängen weitgehend von den je spezifischen Fragestellungen und methodischen Voraussetzungen des theoretischen Rahmens der Untersuchung ab.“ (Bußmann, 2008, S.378)

Korpuslinguistik: „methodischer Ansatz, sprachwissenschaftliche Aussagen auf die (meist statistische) Analyse von Korpora zu stützen. Das Korpus ist einerseits die empirische Quelle der einzelnen Belege, andererseits der Prüfstein für Verallgemeinerungen.“ (Bußmann, 2008, S.379)

Wortform: „konkret realisierte grammatikalische Form eines Wortes im Kontext eines Satzes. Das dem Lexem als der (unveränderlichen) abstrakten Basiseinheit des Lexikons entsprechende Wort der Oberflächenstruktur wird nach grammatikalischen Kategorien (wie Tempus, Numerus, Kasus, Person u.a.) in abgewandelten Wortformen realisiert." (Bußmann, 2008, S.799)

Lemma: „ist die Grundform einer bestimmten lexikalischen Einheit und steht stellvertretend für alle Wortformen dieser lexikalischen Einheit" (Lemnitzer & Zinsmeister, 2010, S.189)

Frequenzangaben:

- Absolute Häufigkeit: „Die absolute Gebrauchshäufigkeit eines Wortes in einem Korpus gibt an, wie oft das Wort insgesamt in allen Texten dieses Korpus vorkommt." (Keibel, 2008, S. 2)
- Relative Häufigkeit: „Die relative Gebrauchshäufigkeit eines Wortes in einem Korpus gibt an, welchen Anteil das Wort an diesem gesamten Korpus ausmacht." (Keibel, 2008, S.2) Dieser Anteil wird als Dezimalzahl, manchmal aber auch in Prozent oder Promille angegeben, vorzugsweise jedoch als Instanzen pro Million Wörter (Keibel, 2008).
- logarithmierte Häufigkeit pro Million: betont die Bedeutung niederfrequenter Wörter stärker als es die relative Häufigkeitsangabe vermag (Gulikers, Rattink, & Piepenbrock, 1995).

Als untere Grenze des gemeinsprachlichen Wortschatzes der deutschen Sprache kann man etwa 200.000 Wörter annehmen, werden Fachausdrücke hinzugenommen, kommt man auf mehrere Millionen Wörter (Best, 2006).

Worthäufigkeit und Wortlänge stehen im direkten Zusammenhang: kurze Wörter werden am häufigsten genutzt – einsilbige Wörter machen fast 50% des 11 Millionen Wörter umfassenden Korpus von Kaeding aus (Best, 2006). Dies hängt laut Best (2006)

auch damit zusammen, dass mit zunehmender Wortlänge die Zahl der Bedeutungen eines Wortes abnimmt.

2.1.1 GESCHICHTE DER WORTHÄUFIGKEITSFORSCHUNG IM DEUTSCHEN SPRACHRAUM

Die erste großangelegte Worthäufigkeitszählung im deutschen Sprachraum wurde 1897 von Friedrich Wilhelm Kaeding veröffentlicht. Er verfolgte damit die Absicht, zur Verbesserung der Stenographie beizutragen. Er erfasste nahezu 11 Millionen (10.910.777) Wörter und nahm eine Auszählung von Wort-, Silben- und Buchstabenhäufigkeiten vor. Kaeding bemühte sich um eine möglichst alle Wissensgebiete umfassende, ausgewogene Zusammenstellung des Korpus. Es flossen juristische, kaufmännische, theologische, militärische, medizinische und geschichtliche Texte ein – erfasst wurden unter anderem wissenschaftliche und belletristische Texte, Zeitungstexte, Privatbriefe und die Bibel (Kaeding, 1897).

Helmut Meier knüpfte mit seinem Werk „Deutsche Sprachstatistik“ (1964) an Kaedings Arbeit an. Darin stütze er sich teilweise auf Kaedings Erkenntnisse, erweiterte diese aber auch um zahlreiche Aspekte, zum Beispiel die statistische Erfassung von Wort- und Satzlängen oder Häufigkeiten grammatikalischer Phänomene (Meier, 1964).

Schließlich startete das Institut für deutsche Sprache (IDS) in Mannheim im Jahr 1964 das Projekt „Mannheimer Korpus 1“, aus welchem sich in weiterer Folge das Deutsche Referenzkorpus (DeReKo) entwickelte (Kupietz & Keibel, 2009).

2.1.2 NEUESTER STAND DER WORTHÄUFIGKEITSFORSCHUNG

Brysbaert und New (2009) betonen die für die Korpuserstellung herangezogenen Quellen als äußerst wichtigen Faktor für die Qualität von Worthäufigkeitsangaben. In

der Vergangenheit wurden Bücher, Zeitungen und Magazine für den Forschungsbereich der visuellen Worterkennung als wichtigste Quellen betrachtet (Brysbaert & New, 2009). Als mit dieser Art von Quellen verbundene Probleme identifizieren Brysbaert und New (2009) die durch Lektorierung künstlich verbesserte Sprache, die - um Wortwiederholungen zu vermeiden – übertriebene lexikalische Variation, und die Tatsache dass die behandelten Themen nicht dem normalen Alltag von durchschnittlichen Menschen entsprechen. In diesem Sinne sehen die oben genannten Autoren das Internet als neue Quelle für spontane und der sprachlichen Realität entsprechenden Verwendung von Worten. So plädierten beispielsweise auch Lund und Burgess (1996) für die Verwendung von user groups als Quelle für sprachliche Informationen und konstruierten auf der Basis von Internet-Newsgroups ein 131 Millionen Wörter umfassendes, allerdings nicht öffentlich zugängliches Korpus - das HAL-Korpus (Hyperspace Analogue to Language). Die Häufigkeitsangaben im HAL-Korpus waren denen des Kucera&Francis-Korpus bei der Vorhersage von Reaktionszeiten bei einer Worterkennungsaufgabe eindeutig überlegen (Burgess & Livesay, 1998).

Blair, Urland und Ma (2002) verglichen die Worthäufigkeitsangaben aus CELEX und dem Kucera&Francis-Korpus mit aus Internetsuchmaschinen gewonnenen Frequenzangaben. Es zeigte sich, dass Letztere eine hohe Validität und Reliabilität aufweisen konnten (Blair, Urland, & Ma, 2002).

New, Brysbaert, Veronis und Pallier (2007) argumentieren, dass die Untertitel von Filmen und TV-Programmen eine gute Grundlage für die Schätzung von Worthäufigkeiten liefern könnten, da sie leicht zugänglich seien und der Alltagssprache eher gerecht würden als beispielsweise literarische Werke. Sie nehmen an, dass Filme und Fernsehprogramme für viele Menschen einen substanziellen Anteil des sprachlichen Inputs darstellen, da sie durchschnittlich 3-4 Stunden täglich vor dem Fernseher verbringen (New, Brysbaert, Veronis, & Pallier, 2007). Im Vergleich zeigte sich, dass die aus Untertiteln gewonnenen Worthäufigkeitsschätzungen von höherer

Qualität waren als Angaben aus herkömmlichen Korpora (New, Brysbaert, Veronis, & Pallier, 2007).

Brysbaert und New (2009) erstellten ein englischsprachiges Korpus aus Untertiteln – das SUBTLEX-us-Korpus. Dieses hat einen Umfang von 51 Millionen Wörtern, davon 16,1 Millionen aus TV-Serien, 14,3 Millionen aus Filmen die vor 1990 veröffentlicht wurden, und 20,6 Millionen Wörter aus Filmen, welche nach 1990 veröffentlicht wurden (Brysbaert & New, 2009). Brysbaert und New (2009) konnten zeigen, dass die aus Untertiteln gewonnenen Worthäufigkeitsangaben die Verarbeitungsgeschwindigkeit von Wörtern besser vorhersagen konnten als die Angaben aus dem Kucera&Francis-Korpus und der englischsprachigen Version von CELEX.

Brysbaert, Buchmeier, Conrad, Jacobs, Bölte und Böhl (2011) konstruierten ein auf Untertiteln basierendes Korpus für den deutschsprachigen Raum – SUBTLEX-de – und verglichen die Aussagekraft von dessen Worthäufigkeitsschätzungen mit anderen Quellen. In diesem Vergleich schnitten die Angaben aus CELEX am schlechtesten ab; bessere Ergebnisse konnten sowohl mit den Angaben aus dem Leipzig-Korpus, als auch mit den aus google books gewonnenen Frequenzen erzielt werden. Als qualitativ am besten erwiesen sich jedoch die Frequenzschätzungen aus dem SUBTLEX-de-Korpus (Brybaert, et al., 2011).

2.1.3 BESONDERHEITEN DER VERTEILUNG VON WORTHÄUFIGKEITEN

Bei Worthäufigkeiten kann grundsätzlich keine Normalverteilung angenommen werden (Dunning, 1993). Sowohl in Frequenzwörterbüchern und großen Korpora als auch in einzelnen Texten folgen die Worthäufigkeiten einer Zipf'schen Verteilung, d.h. es gibt einige wenige hochfrequente Wörter und eine sehr große Anzahl an niederfrequenten Wörtern (Best, 2006). So machen beispielsweise im 9 Millionen Wörter (Grundformen) umfassenden Dudenkorpus die 100 häufigsten Wörter fast die

Hälfte des Gesamtkorpus aus! (Bibliographisches Institut GmbH, 2013)

Werden Wörter aus einem Korpus nach ihrer Häufigkeit geordnet, so gilt r*f=C, d.h. das Produkt aus Rang und Frequenz ergibt eine Konstante (Best, 2006). So ist beispielsweise die Frequenz des Wortes mit Rang 50 dreimal so hoch wie die des Wortes auf Rang 150 (Manning & Schütze, 1999). Best (2006) erklärt, dass dieses Gesetz die mittleren Ränge sehr gut modelliert, bei extrem häufigen und extrem seltenen Wörtern gibt es jedoch Abweichungen. Best (2006) argumentiert, dass die Zipf-Mandelbrot-Verteilung

$$P_{x=\frac{(b+x)^{-a}}{Fn}} \qquad x = 1, 2, \dots, n \qquad F_{(n)=\sum_{i=1}^{n}(b+i)^{-a}}$$

für diese Fälle ein geeignetes Modell darstellt und die Abweichungen in den Randbereichen korrigiert.

2.2 WORD FREQUENCY – FREQUENT WORDS

Die folgenden Abschnitte geben einen kurzen Einblick in die wissenschaftlichen Erkenntnisse zum Worthäufigkeitseffekt. Es werden ausgewählte Erklärungsansätze und auch kritische Beiträge bzw. alternative Erklärungen präsentiert.

2.2.1 ZUR WAHRNEHMUNG VON HÄUFIGEN WÖRTERN

Kommt ein Wort besonders häufig in einer Sprache vor, so erscheint es wenig überraschend, dass es leichter und schneller gelesen/verarbeitet wird als ein seltenes Wort - Erfahrung sollte sich allgemein positiv auf die Performance auswirken (Balota, Yap, & Cortese, 2006).

Obwohl es naheliegend erscheinen mag, wie und warum Worthäufigkeit die Worterkennung beeinflusst, gibt es unterschiedliche theoretische Erklärungsmodelle dazu (Balota, Yap, & Cortese, 2006).

Einige der Erklärungen basieren auf Morton's (1969) Logogen-Modell. Nach diesem Modell haben hochfrequente Wörter aufgrund der höheren Wahrscheinlichkeit ihres Auftretens eine niedrigere Aktivierungsschwelle als niederfrequente Wörter (Cortese & Balota, 2012).

Das Zwei-Wege-Modell (Dual-Route-Cascaded-Model) von Coltheart (2001) geht von einer lexikalischen und einer sublexikalischen Route aus und erklärt den Worthäufigkeitseffekt durch Aktivierungsmuster bei der lexikalischen Route (Balota, Yap, & Cortese, 2006).

Bei den sogenannten Modellen der geordneten Suche (Forster, 1976; Murray & Forster, 2004) wird das mentale Lexikon systematisch durchsucht, wobei hochfrequente vor niederfrequenten Wörtern abgerufen werden (Cortese & Balota, 2012).

2.2.2 DER WORTHÄUFIGKEITSEFFEKT

Die Häufigkeit, mit der ein Wort in einer Sprache vorkommt, beeinflusst nahezu alle Bereiche der Worterkennung (Balota, Yap, & Cortese, 2006). So konnten Worthäufigkeitseffekte bei der lexikalischen Entscheidung, der Wortbenennung, der Wortwahrnehmung und der Fixationsdauer nachgewiesen werden (Balota, Yap, & Cortese, 2006). Balota, Yap und Cortese (2006) stellen fest, dass dieser Umstand wenig überraschend ist, da die Häufigkeit, mit der ein Wort in der Sprache vorkommt, auch die Häufigkeit, mit der man mit einem Wort konfrontiert ist, bedingt, und die Erfahrung somit die Performance beeinflusst.

Solomon und Howes (1951) zeigten in einem Experiment, dass der *visual duration threshold* eines Wortes eine annähernd lineare Funktion des Logarithmus der relativen Häufigkeit des betreffenden Wortes ist.

Schon seit langem ist bekannt, dass die Häufigkeit, mit der ein Wort im Sprachgebrauch vorkommt, einen großen Einfluss auf die Benennungsgeschwindigkeit hat: je häufiger ein Wort auftritt, desto schneller wird es benannt (Spieler & Balota, 2000). Die Worthäufigkeit scheint laut Spieler und Balota (2000) mehrere Stufen des Übersetzungsprozesses von visuellem Stimulus zum phonologischen Output zu beeinflussen. Sie vermuten zwar, dass ein großer Anteil der Häufigkeitseffekte aus der Verarbeitung der Phonologie entspringt, nehmen jedoch an, dass die Häufigkeit Einfluss auf den Zugang zu semantischer Information hat und auf die Initiation und Ausführung des Artikulierungsprogramms, welches für den verbalen Output notwendig ist. In den meisten Modellen visueller Worterkennung spielt Worthäufigkeit eine entscheidende Rolle (Spieler & Balota, 2000).

Duyck, Desmet, Verbeke & Brysbaert (2004) bezeichnen die Worthäufigkeit als eine der wichtigsten linguistischen Variablen: Wörter, die häufiger vorkommen, werden schneller und exakter verarbeitet als solche, die seltener vorkommen. Dieser Effekt wurde erstmals bei der *tachistoscopic recognition* (Howes & Solomon, 1951) demonstriert und später auf ein breiteres Spektrum an Aufgaben generalisiert, darunter lexikalische Entscheidungen (z.B. Whaley, 1978) und Wortbenennung (z.B. Forster & Chambers, 1973). Es ist wichtig, dass Worthäufigkeit in psycholinguistischen Experimenten berücksichtigt wird, da diese Variable subtile Auswirkungen hat, nicht nur zwischen hoch- und niederfrequenten Wörtern, sondern auch zwischen häufigen und nur geringfügig weniger häufigen Wörtern (Duyck, Desmet, Verbeke, & Brysbaert, 2004).

Brysbaert, Buchmeier, Conrad, Jacobs, Bölte und Böhl (2011) betonen in ihrer Arbeit, dass Worthäufigkeit eine der wichtigsten Variablen in der experimentellen Psychologie sei. Sie ist unter anderem der beste Prädiktor für die Geschwindigkeit lexikalischer

Entscheidungen, d.h. für die Geschwindigkeit, mit der eine Buchstabenreihe als Wort oder „non-word“ identifiziert wird (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004).

Kronbichler et al. (2004) konnten zeigen, dass bestimmte Regionen im *left fusiform gyrus*, die mit Bereichen der visuellen Worterkennung korrespondieren, auf Worthäufigkeit reagieren, und zwar mit sinkender Aktivierung bei steigender Worthäufigkeit, d.h. wenn häufige Wörter gelesen werden, ist geringere Aktivierung in den Bereichen der visuellen Worterkennung nötig (Kronbichler, et al., 2004).

2.2.3 KRITIK / ANDERE ERKLÄRUNGEN ZUM WORTHÄUFIGKEITSEFFEKT

Mitte der 1990er-Jahre wurde vermutet, dass alle in der Literatur bekannten Häufigkeitseffekte eigentlich konfundierte *Age-of-Acquisition*-Effekte darstellen (Morrison & Ellis, 1995). Das *Age-of-Acquisition* eines Wortes ist jenes Alter, in dem das Wort erlernt wird. Die Hypothese von Morrison und Ellis (1995) lautete, dass häufige Wörter nicht schneller verarbeitet werden, weil sie häufig sind, sondern weil sie generell früher erworben wurden. Da viele hochfrequente Wörter auch relativ früh im Leben erworben und viele niederfrequente Wörter erst relativ spät erworben werden, gibt es eine natürliche Konfundierung zwischen Worthäufigkeit und *Age-of-Acquisition* (Ghyselinck, Lewis, & Brysbaert, 2004). Zevin und Seidenberg (2002) vermuten, dass die Häufigkeitsangaben in CELEX einen Bias aufweisen, da ausschließlich Literatur für Erwachsene in den Quellen enthalten ist und somit die Häufigkeiten früh erworbener Wörter systematisch unterschätzt werden, und dass dieser Umstand einen Teil des Age-of-Acquisition-Effekts erklären könnte. Gegenwärtig scheint es, als hätten sowohl Häufigkeit als auch *Age-of-Acquisition* voneinander unabhängige Effekte bei der Wortverarbeitung (Brysbaert, Lange, & Van Wijnendaele, 2000). Worthäufigkeit wird nach wie vor in so gut wie allen Studien zur

Wortverarbeitung kontrolliert bzw. manipuliert (Duyck, Desmet, Verbeke, & Brysbaert, 2004).

Das Konstrukt des mentalen Lexikons kann als eines der experimentell und modelltheoretisch am besten untersuchten angesehen werden (Graf, Nagler, & Jacobs, 2005). Als am häufigsten angewandte Methode in diesem Forschungsbereich gilt laut Graf, Nagler und Jacobs (2005) die Erhebung von behavioralen Daten, wie Reaktionslatenzen und Fehlerzahlen im Rahmen von Benennungsaufgaben und lexikalischen Entscheidungsaufgaben. Weitere Zugänge sind beispielsweise elektrophysiologische Methoden und Magnetresonanzmethoden.

Aus zahlreichen Forschungsbeiträgen ergaben sich bis dato mehr als 50 Variablen, welche auf das messbare kognitive Geschehen bei der Wortverarbeitung einen Einfluss ausüben, darunter Wortfrequenz, Wortlänge, Nachbarschaftsverhältnisse und Konsistenz (Graf, Nagler, & Jacobs, 2005). Die unabhängige Manipulation der Variablen im Experiment und die statistischen Analysen werden durch den Umstand erschwert, dass eine Vielzahl dieser Variablen interkorreliert ist. Beispielsweise sind hochfrequente Wörter tendenziell kurz, und kurze Wörter wiederum haben im Vergleich zu langen Wörtern meist mehr orthographische Nachbarn (Graf, Nagler, & Jacobs, 2005). Graf, Nagler und Jacobs (2005) streben in ihrem Beitrag die Reduktion der Komplexität des Zusammenhangs zwischen relevanten Worterkennungsvariablen an, mit dem Ziel, eine bessere Kontrolle von Wörtern in psychologischen Experimenten zu ermöglichen. Sie führten auf Grundlage der 57 Eingangsvariablen eine Faktorenanalyse mit schiefwinkeliger Rotation durch, woraus sich 15 interkorrelierende Faktoren ergaben, welche 81% der Varianz der Eingangsvariablen erklärten. Durch eine zweite Faktorenanalyse mit orthogonaler Rotation ergaben sich sechs orthogonale Faktoren zweiter Ordnung. Faktor 1 wird als „allgemeine Worthaftigkeit“ interpretiert; Wörter mit hohem Faktorwert auf F1 sind typischerweise kurz und verfügen über viele Nachbarn, d.h. sind anderen Wörtern ähnlich.

Balota und Chumbley (1984) untersuchten die Rolle der Worthäufigkeit bei lexikalischen Entscheidungsaufgaben. Bei der Untersuchung von Variablen, die die Geschwindigkeit des lexikalischen Zuganges beeinflussen, wurde dem Vorgang der lexikalischen Entscheidung (*lexical decision task*) viel Beachtung geschenkt (Balota & Chumbley, 1984). Hierbei geht es um das Erkennen, ob eine Zeichenkombination ein Wort ist oder nicht. Ergibt sich hierbei eine Veränderung der Reaktionsgeschwindigkeit durch Veränderung einer Variable, wird meist angenommen, dass diese Variable einen Effekt darauf hat, wie gut oder wie schnell relevante Information über das zu benennende Wort abrufbar ist (= Zugang zu seiner lexikalischen Repräsentation) (Balota & Chumbley, 1984). Dieser Forschungszugang setzt laut Balota und Chumbley (1984) voraus, dass lexikalischer Zugang (*lexical access*) der einzige Prozess im *lexical decision task* ist, der von der manipulierten Variable beeinflusst wird. Sie geben daher zu bedenken, dass die Rolle der Worthäufigkeit dadurch tendenziell überschätzt werden könnte (Balota & Chumbley, 1984).

Brysbaert, Buchmeier et al. (2011) kritisieren die Tatsache, dass angesichts des großen Einflusses der Worthäufigkeit wenige Studien in der Vergangenheit darauf abzielen, die Zuverlässigkeit von Worthäufigkeitsangaben zu überprüfen. So werden beispielsweise im englischsprachigen Raum in den meisten Studien die Häufigkeitsangaben aus dem Korpus von Kucera und Francis (1967) verwendet, obwohl Brysbaert und New (2009) nachweisen konnten, dass diese aufgrund der geringen Korpusgröße und relativ alter Quellen äußerst ungenau zu sein scheinen. Brysbaert, Buchmeier et al. (2011) regen eine sorgfältige Qualitätskontrolle der Häufigkeitsangaben an. In ihrer Studie untersuchten sie Variablen, welche die Qualität von Worthäufigkeitsangaben beeinflussen: sie stellten fest dass die Korpusgröße einen entscheidenden Einfluss hat, vor allem auf die Häufigkeitsangaben niederfrequenter Wörter. Weiters konnten sie den Sprachstil (*language register*) und die Aktualität eines Korpus bzw. seiner Quellen als wichtige Einflussfaktoren auf die Qualität von Frequenzangaben identifizieren.

2.3 KORPUSDATENBANKEN

Die folgenden Kapitel erklären den grundsätzlichen Aufbau von Korpusdatenbanken und geben einen Überblick über große Korpusdatenbanken der deutschen Sprache.

2.3.1 DATENBANK VS LEXIKON

Eine Wortdatenbank unterscheidet sich bereits im Aufbau stark von einem herkömmlichen Lexikon: in einem Lexikon finden sich alphabetisch geordnete Einträge mit verschiedenen grammatikalischen/linguistischen Informationen zu jedem einzelnen Wort - es ermöglicht, jeweils zu einem konkreten Wort gewünschte Informationen abzurufen (Burnage, 1990). In der simpelsten Form kann auch eine Datenbank wie ein Wörterbuch funktionieren: eine Liste von Wörtern mit bestimmten Informationen zu jedem Wort. Der erste wesentliche Unterschied zwischen einer Computerdatenbank und einem Lexikon ist nach Burnage (1995) die Anordnung der Zusatzinformationen in Spalten. Während es in einem Lexikon lediglich möglich ist, die Informationen zu einem konkreten Wort abzurufen, bietet eine Computerdatenbank die Möglichkeit, Informationen selektiv abzurufen.

Der entscheidende Unterschied zwischen einer Datenbank und einem Lexikon ist also die Flexibilität beim Abrufen und der Darstellung der gewünschten Daten (Burnage, 1990). Ein weiterer gravierender Unterschied besteht darin, dass nicht wie im Lexikon nur nach einem konkreten Schlagwort gesucht werden kann, sondern nach beliebigen Kriterien. So ermöglicht CELEX beispielsweise die Erstellung von Listen bestimmter Wortklassen (z.B. nur Verben). Der Benutzer kann so eigene Lexika erstellen, die genau auf seine Bedürfnisse zugeschnitten sind.

2.3.2 KORPUSDATENBANKEN IM DEUTSCHEN SPRACHRAUM

Im deutschen Sprachraum existiert mittlerweile eine relativ große und ständig wachsende Anzahl an linguistischen Korpora. An dieser Stelle werden einige ausgewählte Korpusprojekte kurz vorgestellt, in den darauffolgenden Kapiteln erfolgt eine genauere Betrachtung der in der vorliegenden Arbeit untersuchten Korpusprojekte CELEX und DeReKo.

DeReKo: Das Deutsche Referenzkorpus des Instituts für Deutsche Sprache Mannheim umfasst über 5,4 Milliarden laufende Wörter (Stand Februar 2012), mit einem Zuwachs von über 300 Millionen Wörtern pro Jahr (Institut für Deutsche Sprache, 2012b). Es ist die größte linguistisch motivierte Sammlung deutscher Texte (Kupietz & Keibel, 2009).

CELEX: Die Datenbank des Centre for Lexical Information enthält Korpora in deutscher, englischer und niederländischer Sprache und ermöglicht die Recherche umfangreicher lexikalischer Informationen. Diese sind entweder über eine CD-ROM oder die Internetanwendung WebCelex (http://celex.mpi.nl/) abrufbar. Das deutschsprachige Korpus enthält etwa 6 Millionen laufende Wörter aus geschriebenen und gesprochenen Quellen (Baayen, Piepenbrock, & Gulikers, 1995).

DWDS – Digitales Wörterbuch der Deutschen Sprache: Projekt der Berlin-Brandenburgischen Akademie der Wissenschaften. Ziel des Projekts ist die Schaffung eines umfassenden, jedem Benutzer über das Internet zugänglichen Wortinformationssystems, das Auskunft über den deutschen Wortschatz in Vergangenheit und Gegenwart gibt (Berlin-Brandenburgische Akademie der Wissenschaften, 2012). Das DWDS baut auf dem sechsbändigen Wörterbuch der Deutschen Gegenwartssprache auf und verknüpft es mit eigenen Text- und Wörterbuchressourcen. Das zeitlich und nach Textsorten ausgewogene Kernkorpus umfasst mehr als 100 Millionen laufende Wörter. Das DWDS ist über die Website http://www.dwds.de frei zugänglich, nach einer – aus urheberrechtlichen Gründen erforderlichen - kostenlosen Anmeldung steht ein größeres Korpus zur Verfügung (Universität Regensburg, 2006).

Das Kernkorpus des DWDS bildet auch die Grundlage für das Projekt **dlexDB** (http://www.dlexdb.de/project/description/). Ziel dieses Projektes ist die Erstellung einer lexikalischen Datenbank für die Verwendung in der psychologischen und linguistischen Forschung.

deWaC: (WaC = Web as Corpus) Das deWaC ist ein rein webbasiertes Korpus, welches durch automatisches Webcrawling im Rahmen der WaCky („Web-as-Corpus kool ynitiative“) erstellt wurde (Baroni & Kilgarriff, 2006). In der ersten Fassung enthält das deWaC 1,7 Milliarden Tokens (Belica, Keibel, Kupietz, & Perkuhn, 2007).

Projekt Deutscher Wortschatz: Im Leipziger Projekt Deutscher Wortschatz werden elektronisch vorliegende Textdaten gesammelt und ausgewertet. Das Korpus umfasst mehr als 2 Milliarden laufende Wörter; das daraus erzeugte Wörterbuch ist unter www.wortschatz.uni-leipzig.de online zugänglich (Quasthoff, 2009).

2.4 CELEX

In den folgenden Abschnitten sollen die Entstehung und der Aufbau der Datenbank CELEX näher beschrieben werden.

2.4.1 ENTSTEHUNG UND ZUSAMMENSETZUNG VON CELEX

Die Abkürzung CELEX steht für das Dutch Centre for Lexical Information. Es wurde als gemeinschaftliches Projekt der Universität von Nijmegen, des Instituts für Niederländische Lexikologie in Leiden, des Max-Planck-Instituts für Psycholinguistik in Nijmegen und des Instituts für Wahrnehmungsforschung in Eindhoven realisiert. Heute ist CELEX Teil des Max-Planck-Instituts für Psycholinguistik (Baayen, Piepenbrock, & Gulikers, 1995).

Die lexikalische Datenbank von CELEX enthält Korpora in deutscher, englischer und niederländischer Sprache und ermöglicht die Recherche umfangreicher lexikalischer Informationen, wie etwa Orthographie, Phonologie, Morphologie, Syntax und Wortfrequenz. Diese Informationen sind entweder über eine CD-ROM oder die Internetanwendung WebCelex (http://celex.mpi.nl/) abrufbar (Baayen, Piepenbrock, & Gulikers, 1995).

Quelle für das deutsche Korpus ist das Mannheim Korpus des Instituts für Deutsche Sprache, welches in der Version von 1984 in etwa sechs Millionen Wörter aus geschriebenen und auch einigen gesprochenen Quellen enthält. Bei den schriftlichen Quellen umfasst die Bandbreite von gehobener bis anspruchsloser Literatur, wissenschaftlicher und populärwissenschaftlicher Literatur, Memoiren, Zeitungen und Magazine. Die gesprochenen Quellen beinhalten Transkriptionen spontaner Sprache, d.h. die Sätze lagen nicht in schriftlicher Form vor, bevor sie in Konversationen, Diskussionen oder Reden benutzt wurden. (Gulikers, Rattink, & Piepenbrock, 1995)

Das CELEX-Korpus besteht aus 5,4 Millionen deutschen Tokens aus geschriebenem Text wie etwa Zeitungen, Romanen und Sachtexten, und 600.000 Tokens aus transkribierter Sprache. Ersteres setzt sich zusammen aus dem „Mannheim Korpus I“, dem „Mannheim Korpus II“ und dem „Bonner Zeitungskorpus I“; letzteres ist aus dem „Freiburger Korpus“ entnommen. Alle genannten Korpora sind auch einzeln über die Software COSMASII des Instituts für Deutsche Sprache in Mannheim abrufbar (Baayen, Piepenbrock, & Gulikers, 1995).

Das Korpus ist relativ unausgewogen, einerseits aufgrund seiner geringen Größe, andererseits durch die Auswahl der darin aufgenommenen Texte. Alle im Korpus enthaltenen Texte wurden zwischen 1949 und 1975 veröffentlicht, dies legt die Vermutung nahe, dass sich die Häufigkeitsangeben in CELEX auf einen veralteten Wortschatz beziehen.

Mannheim Korpus I: besteht aus 293 Texten aus den Jahren 1950 bis 1967 und umfasst 2.200.000 laufende Wortformen. Es enthält eine Auswahl belletristischer

Texte (Heinrich Böll: Ansichten eines Clowns; Werner Bergengruen: Das Tempelchen; Max Frisch: Homo Faber; Günter Grass: Die Blechtrommel; Uwe Johnson: Das dritte Buch über Achim; Thomas Mann: Die Betrogene; Erwin Strittmatter: Ole Bienkopp), die Memoiren von Theodor Heuss („Erinnerungen 1905-1933“), wissenschaftliche und populärwissenschaftliche Literatur, Trivialliteratur und Artikel aus Zeitungen und Zeitschriften (Institut für Deutsche Sprache, 2012b).

Mannheim Korpus II: umfasst 52 Texte mit ca. 300.000 laufenden Wortformen. Die Texte stammen aus den Jahren 1949 bis 1952 und 1960 bis 1974. Enthalten sind Erlasse, Satzungen, Beschlüsse, Gebrauchsanweisungen, Lehrbücher, Nachrichten, Prospekte, Trivialliteratur, wissenschaftliche und populärwissenschaftliche Literatur und Artikel aus Zeitungen und Zeitschriften (Institut für Deutsche Sprache, 2012b).

Bonner Zeitungskorpus: enthält 10.840 Texte mit etwa 3.100.000 laufenden Wortformen. Es umfasst Jahrgangsquerschnitte (1949, 1954, 1959, 1964, 1969 und 1974) aus Artikeln der Tageszeitungen „Neues Deutschland“ (DDR) und „Die Welt“ (BRD) (Institut für Deutsche Sprache, 2012b).

Freiburger Korpus: enthält 222 Tonaufnahmen und 221 Transkripte mit einer Gesamtdauer von 68 Stunden. Die Texte wurden in den Jahren 1960 bis 1974 aufgenommen. Sie setzen sich zusammen aus Interviews, Diskussionen, Unterhaltungen, Vorträgen, Reportagen und Erzählungen, und wurden großteils von Rundfunk- und Fernsehanstalten beschafft. Es wurden Fernseh- und Rundfunkaufnahmen mitgeschnitten und darüber hinaus auch eigene Aufnahmen aus privaten und öffentlichen Kommunikationssituationen gemacht (Institut für Deutsche Sprache, 2004).

2.4.2 AUFBAU UND BENUTZUNG VON CELEX

CELEX stellt verschiedene Lexika in den Sprachen Englisch, Deutsch und Niederländisch zur Verfügung. Für die vorliegende Arbeit wurde lediglich der deutschsprachige Teil der

Datenbank verwendet. Eine weitere Unterteilung erfolgt in Lemma, Wortform und *Mannheim Corpus Types* (Burnage, 1990).

Wortformen können als „echte Wörter“ betrachtet werden, so wie sie in der Sprache tatsächlich gebraucht werden. Ein Wortformlexikon zeigt also nicht wie ein Wörterbuch abstrakte Nennformen, die eine bestimmte Wortgruppe repräsentieren, sondern listet wirklich jedes im Sprachgebrauch mögliche Wort auf (Burnage, 1990). Burnage (1990) führt hier als Beispiel an, dass während im Lemma-Lexikon lediglich das Wort *Kind* angezeigt wird, das Wortformlexikon alle gebräuchlichen Abwandlungen wie etwa *Kind*, *Kindes*, *Kinde*, *Kinder* und *Kindern* enthält. Zusätzlich kann auch die Information zum jeweiligen Lemma bzw. Wortstamm angezeigt werden.

Für Wortformen bietet CELEX umfangreiche orthographische, phonetische und morphologische Informationen, sowie Informationen über die Häufigkeit (Burnage, 1990).

Häufigkeitsinformationen liefert CELEX laut Burnages „CELEX – a guide for users“ (1990) zu Lemmata, Wortformen und den sogenannten *Mannheim Corpus Types*: Mannheim tokens sind Strings im Mannheim Textkorpus, welche mindestens ein alphabetisches Zeichen zusammen mit null oder mehreren alphanumerischen Zeichen enthalten, z.B. *fünfzehn* oder *15jährige*. In der *Mannheim Corpus Types*-Liste sind alle Types erfasst, die in mindestens zwei verschiedenen Korpustexten vorkommen (Burnage, 1990).

2.4.3 WORTHÄUFIGKEITSANGABEN IN CELEX

Im „German Linguistic Guide“ liefern Gulikers, Rattink und Piepenbrock (1995) eine ausführliche Beschreibung der in CELEX verfügbaren Kennzahlen. Die Häufigkeitsinformationen in CELEX sind sowohl für Lemmata als auch für Wortformen verfügbar (Gulikers, Rattink, & Piepenbrock, 1995).

Gulikers et al. (1995) erklären, dass der Ausgangspunkt der Berechnung der Worthäufigkeiten das sechs Millionen Wörter umfassende Mannheim Korpus ist. Es wird gezählt, wie oft jeder einzelne String auftritt. Daraus ergeben sich die *raw string counts* – sie geben an, wie oft jede Zeichenkombination im Korpus vorkommt. Unterschiedliche Bedeutungen oder Wortklassen werden hierbei nicht berücksichtigt. Um aus diesen Angaben zu aussagekräftigeren Zahlen zu kommen, werden im nächsten Schritt zu bestimmten Lemmata gehörige Wortformen identifiziert und von anderen Strings wie Eigennamen oder Fremdwörtern unterschieden. Diese Unterscheidung ist in manchen Fällen eindeutig – so ist der String *Bezirken* immer die Dativ-Plural-Form des Lemmas *Bezirk*. In diesem Fall ist die *raw-string-frequency* des Strings *Bezirken* auch die Häufigkeit der Wortform *Bezirken* (Gulikers, Rattink, & Piepenbrock, 1995).

Sobald die Häufigkeiten für die Wortformen feststehen, ergibt sich daraus auch die Häufigkeit der Lemmata – die Häufigkeiten der Wortformen werden addiert. So setzt sich die Häufigkeit des Lemmas *Bezirk* zusammen aus den Häufigkeiten der Nominativ-, Dativ- und Akkusativform Singular der Wortform *Bezirk*, plus der Genitiv-Form *Bezirks*, der Häufigkeit der Nominativ-, Genitiv- und Akkusativform Plural *Bezirke* und der Dativform Plural *Bezirken* (Gulikers, Rattink, & Piepenbrock, 1995).

2.4.3.1 HÄUFIGKEITSANGABEN BEI HOMOGRAPHISCHEN WÖRTERN

Schwieriger gestaltet sich laut Gulikers et al. (1995) die Auswertung der Häufigkeiten homographischer Wortformen, d.h. Wortformen mit gleicher Schreibweise aber unterschiedlicher Bedeutung. Die Autoren schlagen in ihrem „German Linguistic Guide“ eine neue Methode zur Unterscheidung der Häufigkeiten homographischer Wörter von, diese wurde aber bis jetzt noch nicht umgesetzt.

Gulikers et al. (1995) erklären, dass um die Häufigkeiten homographischer Wörter exakt festzustellen, eine manuelle Auswertung erfolgen müsste, was aufgrund der

Datenmenge praktisch nicht durchführbar ist. Um die exakte Häufigkeit der unterschiedlichen Varianten homographischer Wörter festzustellen, müsste jedes einzelne Auftreten im Korpus manuell überprüft werden. Die Autoren beschreiben, dass es zwar grundsätzlich möglich ist, diese Unterscheidung mittels eines Computerprogrammes vorzunehmen, die Ergebnisse aber nicht von zufriedenstellender Qualität sind. Aus diesem Grund wurde für CELEX beschlossen, diese sogenannte Disambiguierung zukünftig manuell vorzunehmen. Jedes Auftreten eines homographischen Wortes im Korpus wird im Kontext überprüft, um das Wort einem bestimmten Lemma zuzuordnen. Obwohl der Ansatz kosten- und zeitintensiv ist, könnten so zuverlässige Ergebnisse erzielt werden (Gulikers, Rattink, & Piepenbrock, 1995).

Gulikers et al. (1995) erläutern diesen geplanten Vorgang der Disambiguierung an folgendem Beispiel: Für das Wort *Messer* ergibt die manuelle Auswertung 84 Treffer mit der Bedeutung *Schneidewerkzeug*, und keinen mit der Bedeutung *jemand/etwas der/das misst*; Fälle, in denen *Messer* als Personenname verwendet wird, werden in der Häufigkeitszählung nicht berücksichtigt. Sobald die Häufigkeiten der Wortformen klar sind, ergeben sich daraus die Häufigkeiten der Lemmata. So beinhaltet die Häufigkeit des Lemmas *Messer* (Schneidewerkzeug) = 99 auch die Häufigkeiten der Wortformen *Messern* (10) und *Messers* (5), während die Häufigkeit des Lemmas *Messer* (jemand der bzw. etwas, das misst) null beträgt.

Im „German Linguistic Guide“ wird erläutert, dass sich diese Auswertung bei sehr häufigen Strings äußerst aufwändig gestaltet. Die Autoren erklären, dass dieser Aufwand nicht unbedingt notwendig sei, da eine intelligente Schätzung und eine Angabe über die Exaktheit der Schätzung im Normalfall ausreichen sollten: Taucht ein homographisches Wort mehr als 100 Mal im Korpus auf, werden nicht mehr alle einzelnen Vorkommen überprüft, sondern es werden stattdessen nach Zufallsprinzip 100 Auftritte des Strings ausgewählt und analysiert. Auf diese Art kann ein Verhältnis errechnet werden, das den Anteil der jeweiligen Bedeutungen im gesamten Korpus abschätzt (Gulikers, Rattink, & Piepenbrock, 1995).

Folgendes Beispiel von Gulikers et al. (1995) erläutert diesen Vorgang: so tritt beispielsweise der String *nahe* im gesamten Korpus 403 Mal auf. Es kann entweder ein Adjektiv sein, eine Präposition, der Indikativ der ersten Person Singular des Verbs *nahen*, der Konjunktiv der ersten oder dritten Person Singular des Verbs *nahen*, oder der Imperativ Singular. Um die Häufigkeiten zu berechnen, wurden 100 Vorkommen zufällig ausgewählt und manuell disambiguiert. 62 Vorkommen wurden dem Präpositions-Lemma zugeordnet, 38 dem Lemma des Adjektivs, und keines dem Lemma des Verbs nahen. Um die wahre Häufigkeit der zum Lemma des Adjektivs gehörigen Wortform abzuschätzen, wird die Auftrittshäufigkeit in der Stichprobe durch die Größe der Stichprobe dividiert, und danach mit der Gesamthäufigkeit multipliziert: 38/100*403=153. Nach dem gleichen Prinzip ergibt sich für die Häufigkeit der Präposition 250, und null für die Häufigkeit des Verbs.

Die so vorgenommenen Schätzungen sind laut Gulikers et al. (1995) meist hinreichend akkurat, können aber durch die statistische Berechnung von erwarteten Abweichungen noch verbessert werden. Mit folgender Formel wird die Abweichung berechnet:

$$N \times 1.96 \times \sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}$$

N: Gesamthäufigkeit des Strings
n: Anzahl der disambiguierten Items aus der Zufallsstichprobe
p: errechnete Verhältniszahl des zu einem bestimmten Lemma gehörenden Items

Am Beispiel des Adjektivs *nahe* ergibt sich so N=403, n=100, und p=0.38. Daraus ergibt sich eine Abweichung von 33,29. So liegt die wahre Häufigkeit des Adjektivs *nahe* mit zumindest 95%iger Wahrscheinlichkeit zwischen 120 und 186 (Gulikers, Rattink, & Piepenbrock, 1995).

Die Abweichungen scheinen in CELEX in der Spalte MannDev auf. Ist der Wert MannDev höher als die Häufigkeit selbst, kann angenommen werden, dass eine willkürliche Schätzung vorgenommen wurde (Gulikers, Rattink, & Piepenbrock, 1995).

2.4.3.2 HÄUFIGKEITSANGABEN FÜR WORTFORMEN IN CELEX

CELEX stellt unterschiedliche Häufigkeitsangaben für Lemmata und Wortformen zur Verfügung. Die Spalten Mann bzw. MannLemma geben die absolute Häufigkeit im Korpus an. Die Werte in den Spalten MannDev bzw. MannDevLemma geben an, wie akkurat die Häufigkeitsangaben für das jeweilige Wort sind. Wurde ein Wort vollständig disambiguiert, beträgt der Wert 0. Werte über 0 weisen darauf hin, dass eine Schätzung notwendig war. Ist der Wert gleich oder größer als die Häufigkeitsangabe, war eine vollständige Disambiguierung nicht möglich (siehe Kapitel 2.4.3.1). Für bestimmte Fragestellungen, beispielsweise Vergleiche mit Häufigkeitsangaben aus anderen Quellen, können sich die absoluten Häufigkeiten als unpraktisch erweisen. Die Spalten MannMln bzw. MannMlnLemma geben die Häufigkeit pro Million an. Dies erleichtert auf den ersten Blick die Interpretation, geht aber gleichzeitig mit einem – wenn auch geringen - Genauigkeitsverlust einher. So hat etwa das Lemma *beraten* eine absolute Häufigkeit von 503, und das Lemma *Kritik* die Häufigkeit 507; in der Spalte MannMlnLemma erhalten beide den Wert 85. Des Weiteren bietet CELEX auch noch die Spalten MannLog bzw. MannLogLemma an, diese enthalten die logarithmierten Häufigkeiten pro Million. Die logarithmierten Werte heben die Bedeutung von niederfrequenten Wörtern hervor. So wird der Unterschied zwischen Wörtern mit der Häufigkeit 1 und 2 größer (0,30103) als der Unterschied zwischen höherfrequenten Wörtern mit den Häufigkeiten 2001 und 2002 (0,000217). Dies bestätigt mathematisch, was sich intuitiv ergibt: da es viel mehr niederfrequente Wörter gibt, sind Unterschiede in diesem Bereich signifikanter als bei hochfrequenten Wörtern. Bei einem hochfrequenten Wort fällt ein Unterschied von 1 oder 2 kaum ins Gewicht, bei niederfrequenten Wörtern ist er von größerer Bedeutung. Die

logarithmierten Werte liegen zwischen 0 und 5. Bei Wörtern mit der Häufigkeit 0 (pro Million) beträgt auch der logarithmierte Wert 0. Dies ist zwar mathematisch nicht korrekt – $\log_{(x)}0$ existiert nicht – in diesem Kontext aber eher unwichtig: Wörter mit MannLog=0 bzw. MannLogLemma=0 kommen höchstens 8 Mal im gesamten Korpus vor. Nur Wörter mit einer Häufigkeit von 2 oder mehr pro Million bzw. einer Häufigkeit von 9 oder mehr im gesamten Korpus haben einen logarithmierten Wert größer als 0. (Gulikers, Rattink, & Piepenbrock, 1995)

2.4.3.3 HÄUFIGKEITSANGABEN AUS GESCHRIEBENEN UND GESPROCHENEN QUELLEN

Etwa 5.400.000 Wörter im Mannheim Korpus stammen aus schriftlichen Quellen, die restlichen 600.000 kommen aus gesprochenen Texten. Bei Bedarf sind Häufigkeitsinformationen auch getrennt verfügbar. (Gulikers, Rattink, & Piepenbrock, 1995)

2.5 DAS DEUTSCHE REFERENZKORPUS (DEREKO)

In den nachfolgenden Abschnitten werden die Entstehungsgeschichte und der Aufbau des Deutschen Referenzkorpus (DeReKo) beschrieben. Es wird auf die Quellen für das Korpus eingegangen, sowie auf die Benutzung mittels der Recherchesoftware COSMASII. Ferner wird auch der Umgang mit Problemen des Urheberrechts erläutert.

2.5.1 ENTSTEHUNGSGESCHICHTE DES DEREKO

Paul Grebe und Ulrich Engel starteten 1964 am Institut für Deutsche Sprache (IDS) das Projekt „Mannheimer Korpus 1“ und erstellten bis 1967 ein Korpus mit 2,2 Millionen

laufenden Wörtern (Kupietz & Keibel, 2009). Seitdem wurde das Korpus laufend erweitert. Nach Kupietz, Belica, Keibel und Witt (2010) ist das Deutsche Referenzkorpus (DeReKo) heute weltweit eine der Hauptquellen für das Studium der deutschen Sprache. Es ist die größte linguistisch motivierte Sammlung deutscher Texte, bestehend aus Fiktion, wissenschaftlicher Literatur, Zeitungstexten und zahlreichen anderen Textarten (Kupietz & Keibel, 2009). Es werden ausschließlich vollständige, unveränderte und lizensierte Texte verwendet. Hauptzweck ist es, als Basis für wissenschaftliche Studien der deutschen Gegenwartssprache zu dienen (Kupietz & Keibel, 2009).

Kupietz, Belica et al. (2010) erklären, dass das DeReKo im Gegensatz zu anderen großen Korpora wie dem British National Corpus oder dem Digitalen Wörterbuch der Deutschen Sprache (DWDS) keinen Anspruch auf Ausgewogenheit erhebt und diese auch nicht primär anstrebt, da Ausgewogenheit oder Repräsentativität nur auf eine Gesamtpopulation bezogen definiert werden kann. *„The resource itself should not dictate a specific population, nor should it define which properties of the population are of particular relevance"* (Kupietz & Keibel, 2009, S.56). Die Frage der Repräsentativität wird also dem Benutzer überlassen, in Abhängigkeit von der jeweiligen Forschungsfrage. Das DeReKo versteht sich als Urstichprobe, aus der nach Bedarf Auszüge verwendet und Substichproben – sogenannte virtuelle Korpora – zusammengestellt werden können (Kupietz & Keibel, 2009).

Derzeit umfasst das DeReKo über 5,4 Milliarden laufende Wörter (Stand Februar 2012), mit einem Zuwachs von über 300 Millionen Wörtern pro Jahr (Institut für Deutsche Sprache, 2012b).

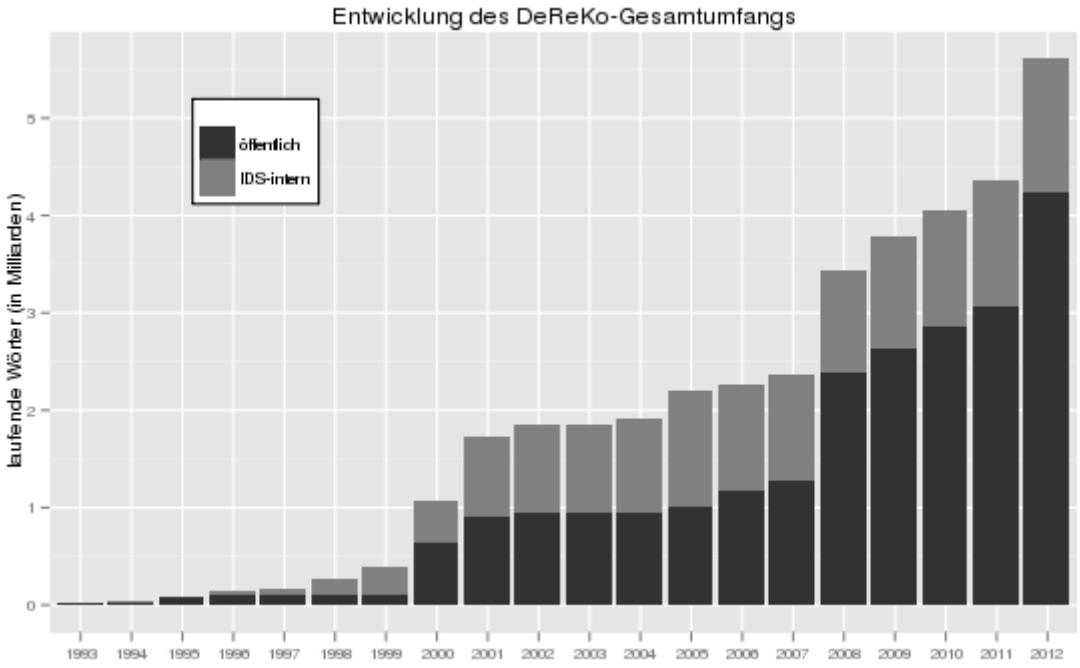

Abbildung 1: jährlicher Zuwachs des DeReKo, Quelle: Institut für Deutsche Sprache (2012b)

2.5.2 QUELLEN DES DEREKO

Das über COSMASII zugängliche DeReKo-Archiv setzt sich aus zahlreichen Teilkorpora zusammen, welche in 9 separaten Archiven organisiert sind (Bopp, 2010).

Bopp (2010) liefert eine Übersicht über die Zusammensetzung des Gesamtkorpus:

- Zeitungs- und Magazintexte: [Deutschland:] Berliner Morgenpost, Bonner Zeitungskorpus, Braunschweiger Zeitung, Computer Zeitung, Frankfurter Rundschau, Hamburger Morgenpost, Hannoversche Allgemeine, Lufthansa Bordbuch, Mannheimer Morgen, Nürnberger Nachrichten, Nürnberger Zeitung, Rhein-Zeitung, spektrumdirekt, VDI Nachrichten, [Österreich:] Burgenländische Volkszeitung, Die Presse, Kleine Zeitung, Neue Kronen-Zeitung, Niederösterreichische Nachrichten, Oberösterreichische Nachrichten, Salzburger Nachrichten, Tiroler Tageszeitung, Vorarlberger Nachrichten, [Schweiz:] Die Südostschweiz, St. Galler Tagblatt, Zürcher Tagesanzeiger

- Historische, v.a. literarische Texte: [18. und 19. Jahrhundert:] Goethe, Brüder Grimm, Marx/Engels; [20. Jahrhundert:] Martin Walser, Mannheimer Korpus 1 (enthält u.a. Heinrich Böll: Ansichten eines Clowns (1963), Max Frisch: Homo Faber (1957), Günter Grass: Die Blechtrommel (1962))
- Gebrauchstexte: Mannheimer Korpus 2 (enthält u.a. Erlasse [1949, 1952, 1960-1974], Gebrauchsanleitungen [1960-1974], Anweisungen [1961, 1966, 1972], Informationsmaterial [1960-1974]), Wikipedia
- Bestehende Korpora: Handbuch-Korpora des IDS, LIMAS-Korpus, Mannheimer Korpus Teile 1 + 2, Wendekorpus
- Morphosyntaktisch annotierte Texte: LIMAS-Korpus, Mannheimer Morgen

Für die vorliegende Diplomarbeit wurde das mehr als 3 Milliarden Wörter umfassende Hauptarchiv „W – Archiv der geschriebenen Korpora“ in der Version von 2012 zur Recherche herangezogen. Dieses enthält Texte aus dem Zeitraum des 18. Jahrhunderts bis zur Gegenwart.

2.5.3 BENUTZUNG DES DEREKO

Da das DeReKo aus lizenzrechtlichen Gründen nicht zum Download zur Verfügung steht, bietet das IDS mit COSMASII eine Software an, welche Nutzern den Zugang zu den im DeReKo enthaltenen Informationen ermöglicht. Die Software erlaubt die Komposition virtueller Korpora, ermöglicht komplexe Suchanfragen – beispielsweise Lemmatisierung, proximity operators, search across sentence boundaries und logische Operatoren – und komplexe Analysen und bietet verschiedene Ansichten für Suchergebnisse (Kupietz, Schonefeld, & Witt, 2010). Laut Kupietz, Schonefeld und Witt (2010) wird aufgrund des enormen Wachstums des DeReKo derzeit an einer neuen Software gearbeitet, da COSMASII bereits an seine Grenzen stößt.

COSMASII (Institut für Deutsche Sprache, 1991-2010) ermöglicht dem Benutzer das DeReKo nach linguistischen Kriterien zu durchsuchen, es erlaubt die Bildung

unterschiedlicher Suchanfragen (Wörter, Teilwörter, Grundformen, Wortklassen, grammatikalische Muster etc.), Darstellung der Ergebnisse nach individuellen Kriterien (Entstehungszeit, Erscheinungsland, Thematik) und Dokumentation und Export der zugehörigen Belege (Bopp, 2010).

2.5.4 PROBLEM DES URHEBERRECHTS

Da viele Texte urheberrechtlich geschützt sind, ist im DeReKo der Zugriff auf vollständige Korpustexte nicht möglich. Das IDS ist bemüht, Vereinbarungen mit Textgebern zu treffen, welche eine öffentliche Nutzung für möglichst viele Teilkorpora ermöglichen. Auf der Website des Instituts für Deutsche Sprache wird der Umgang mit urheberrechtlichen Problemen wie folgt erklärt:

> Durch juristische Vereinbarungen mit Verlagen, Zeitungsredaktionen und Autoren war und ist das IDS in der Lage, urheberrechtlich abgesichertes Textmaterial derart zu beschaffen, dass alle Korpora IDS-intern und Teile dieser Korpora weltweit öffentlich genutzt werden können, und zwar ausschließlich zu wissenschaftlichen, nichtkommerziellen Zwecken. Die Textkorpora des IDS sind zudem ausschließlich über das COSMAS-System recherchierbar; kein Nutzer hat Zugriff auf vollständige Korpustexte, sondern nur auf begrenzte Kontexte zu Suchanfragen. Aufgrund des auch im Hinblick auf eine wissenschaftliche Nutzung strikten Urheberrechts und der daraus resultierenden strengen Vereinbarungen mit unseren Textgebern, die, wenn es sich z.B. um Verlage handelt, selbst wiederum urheberrechtlich und vertraglich gebunden sind, können selbst IDS-intern keine vollständigen Korpustexte zur Verfügung gestellt werden.
>
> In der Vergangenheit hat das Projekt DEREKO I (1999-2001) viel zur Klärung gerade der urheberrechtlichen Unsicherheit beigetragen. Der Anteil der Korpora, die öffentlich zugänglich sind, stieg durch dieses Projekt beträchtlich.

> Seit Mitte 2004 läuft nun eine erneute Akquisitionsinitiative, die verstärkt darauf abzielt, langfristige Vereinbarungen mit den Textgebern zu schließen und möglichst alle neu akquirierten Texte auch IDS-extern nutzbar zu machen. (Institut für Deutsche Sprache, 2012b)

2.5.5 WORTHÄUFIGKEITEN IM DEREKO

Es liegen keine der Autorin bekannten schriftlichen Quellen zur genauen Vorgehensweise bei der Ermittlung der Worthäufigkeiten im DeReKo vor.

3 EMPIRISCHER TEIL

In den folgenden Kapiteln werden das Ziel der Untersuchung sowie das Untersuchungsdesign erläutert. Es erfolgt eine genaue Beschreibung der Datenerhebung und der verwendeten Stichprobe. Des Weiteren folgt die Begründung für die Wahl der Log-Likelihood-Ratio als geeignetes Verfahren für den Vergleich von Wortfrequenzangaben, sowie eine Erklärung der Berechnung der Log-Likelihood-Werte. Im Anschluss werden die Ergebnisse präsentiert.

3.1 UNTERSUCHUNGSZIEL UND HYPOTHESEN

Wie u.a. Brysbaert und New (2009) betonen, sind zuverlässige Angaben von Worthäufigkeiten essenziell, um den Worthäufigkeitseffekt zu untersuchen bzw. als mögliche Stör-/Moderatorvariable zu kontrollieren.

Die vorliegende Arbeit versucht anhand eines Vergleichs der Häufigkeitsangaben ausgewählter Wörter in CELEX und DeReKo festzustellen, ob sich die Worthäufigkeiten in den beiden Korpora signifikant unterscheiden.

Ferner soll untersucht werden, ob es eine systematische Abweichung der Häufigkeitsangaben bei besonders häufigen oder besonders seltenen Wörtern gibt, und ob es gegebenenfalls zu einer systematischen Über- bzw. Unterschätzung der Frequenzangaben kommt. Im Falle von Abweichungen der Häufigkeitsangaben ist ferner auch von Interesse wie stark diese sind.

3.2 UNTERSUCHUNGSDESIGN

Im ersten Schritt wird die gesamte Wortliste aus CELEX bereinigt und anhand der logarithmierten Häufigkeiten pro Million in 4 Gruppen mit unterschiedlicher Häufigkeit geteilt. Aus jeder dieser Gruppen wird eine Zufallsstichprobe gezogen. Im nächsten

Schritt werden mit Hilfe der Recherchesoftware COSMASII die korrespondierenden Häufigkeitsangaben aus dem DeReKo ermittelt. Nun soll für jedes einzelne Wort die Log-Likelihood-Ratio ermittelt werden, welche Auskunft darüber gibt, ob und wie stark die Häufigkeitsangaben aus den beiden Datenbanken voneinander abweichen.

In den folgenden Abschnitten wird näher auf die Auswahl und Zusammensetzung der Stichprobe sowie auf die Berechnung der Log-Likelihood-Werte eingegangen.

3.2.1 TECHNISCHES

Die Worthäufigkeiten aus CELEX wurden der zugehörigen CD-ROM (Baayen, Piepenbrock, & Gulikers, 1995) entnommen.

Die Worthäufigkeitsangaben aus dem DeReKo wurden mit Hilfe der vom IDS Mannheim frei zur Verfügung gestellten Software COSMASII = Corpus Search, Management and Analysis System (Institut für Deutsche Sprache, 1991-2010) ermittelt.

Die statistische Auswertung der Daten und Berechnung der Log-Likelihood-Werte wurden mit SPSS 19 vorgenommen. Die Graphiken und Tabellen wurden mit SPSS 19 und Microsoft Excel erstellt.

3.3 DIE STICHPROBE

Da das DeReKo aus datenschutzrechtlichen Gründen nicht in vollem Umfang verfügbar ist, erfolgt der Vergleich der Häufigkeitsangaben anhand einer aus CELEX gezogenen Zufallsstichprobe. In einem ersten Schritt wurde die vollständige Liste aller Wortformen (N=365.530) aus CELEX bereinigt: doppelte Fälle wurden eliminiert. Die verbleibenden Wortformen (N=359.569) wurden nach Häufigkeiten sortiert und Wörter mit einer log-frequency (logarithmierte Häufigkeit pro Million) von 0 eliminiert,

da sie unbrauchbar für Berechnung der Log-Likelihood-Ratio sind. Eine logarithmierte Häufigkeit von 0 gibt an, dass das Wort weniger als 2 Mal pro Million im Korpus auftritt, es handelt sich somit um extrem seltene Wörter. Nun wurden die Wortformen (N=25.745) anhand der Häufigkeit in 4 Gruppen unterteilt: dazu wurden wieder die logarithmierten Häufigkeiten als Trennscores herangezogen. Wörter mit einer log-frequency von weniger als 1 wurden als sehr selten eingestuft – log-frequency zwischen null und 1 bedeutet, dass das betreffende Wort im Korpus weniger als 10 Mal pro Million Wörter auftritt. Als eher selten wurden Wörter mit einer log-frequency von weniger als 2 eingestuft, sie kommen weniger als 100 Mal pro Million vor. Eine log-frequency zwischen 2 und 2,999 bedeutet, dass das betreffende Wort zwischen 100 und 1.000 Mal pro Million auftritt; diese Wörter wurden als relativ häufig betrachtet. Eine log-frequency von 3 oder mehr schließlich zeigt an, dass das Wort öfter als 1.000 Mal pro Million vorkommt - bei einem Wert von 4 oder mehr sogar über 10.000 Mal – es handelt sich hierbei also um sehr häufige Wörter.

Die Gruppe der häufigsten Wörter – log-frequency größer oder gleich 3 – erwies sich mit n=60 als zu klein, um daraus noch eine Zufallsstichprobe zu ziehen und wurde deshalb vollständig in die Stichprobe aufgenommen. Aus den verbleibenden 3 Gruppen wurde unter Zuhilfenahme des in SPSS verfügbaren Zufallsgenerators Mersenne Twister jeweils eine Zufallsstichprobe im Umfang von n=75 gezogen.

Tabelle 1: Untergruppen der Stichprobe

Gruppen	Grundgesamtheit in CELEX		Gruppe in Stichprobe	
	N	in %	n	in %
Sehr häufig: MannLog 3 oder höher	60	0,23	60	21,1
Häufig: MannLog kleiner als 3	751	2,92	75	26,3
Selten: MannLog kleiner als 2	6.351	24,67	75	26,3
Sehr selten: MannLog kleiner als 1	18.583	72,18	75	26,3
Gesamt	25.745	100,00	285	100,0

Für die so entstandene Gesamtstichprobe wurden nun die korrespondierenden Häufigkeitsangaben aus dem Deutschen Referenzkorpus ermittelt. Die frei erhältliche Software COSMAS II ermöglicht unter anderem die einzelne Abfrage der absoluten Häufigkeit der jeweils interessierenden Wortform.

3.4 WAHL DES VERFAHRENS - LOG-LIKELIHOOD-RATIO

Rayson und Garside (2000) beschreiben die Log-Likelihood-Ratio als angemessenes Verfahren zum Vergleich von Häufigkeitsangaben in unterschiedlichen Korpora. Kilgarriff (1996) kritisiert die Verwendung von Chi-Quadrat-Tests bei großen, nicht randomisierten Stichproben und empfiehlt den Mann-Whitney-Test, welcher Rangdaten vergleicht; durch dessen niedrigeres Skalenniveau geht allerdings wertvolle Information verloren. Dunning (1993) stellt fest, dass Normalverteilung bei statistischen Textanalysen grundsätzlich nicht angenommen werden kann. Die Voraussetzungen, auf welchen Tests wie etwa Pearsons Chi-Quadrat oder z-scores basieren, sind bei statistischen Textanalysen nicht gegeben, da Texte größtenteils aus „seltenen Ereignissen“ bestehen und somit von der Annahme der Normalverteilung sehr weit entfernt sind (Dunning, 1993). Dunning (1993) erklärt, dass während die meisten Wörter in einem laufenden Text häufig sind, die meisten Wörter im Gesamtvokabular eher selten sind – die meisten in der Linguistik angewandten statistischen Verfahren sind jedoch für die Analyse geläufiger, also normalverteilter Ereignisse konzipiert. Der Chi-Quadrat-Test verliert an Reliabilität, wenn die erwartete Häufigkeit kleiner als 5 ist, und überschätzt vermutlich hochfrequente Wörter, deshalb empfehlen Dunning (1993) und Rayson und Garside (2000) als Alternative zu Pearsons Chi-Quadrat die Log-Likelihood-Ratio.

Manning und Schütze (1999) nennen als zusätzlichen Vorteil der Log-Likelihood-Ratio gegenüber Pearsons Chi-Quadrat die einfachere Interpretierbarkeit: Die Log-

Likelihood-Ratio misst die Diskrepanz zwischen den beobachteten Worthäufigkeiten und jenen Werten, die zu erwarten wären, wenn die Häufigkeitsangaben in beiden Texten bzw. Korpora übereinstimmen. Je größer die Diskrepanz, desto größer die Log-Likelihood, und umso größer die statistische Signifikanz der Unterschiede zwischen den Worthäufigkeiten in den Korpora. Einfach ausgedrückt, zeigt der Log-Likelihood-Wert wie viel wahrscheinlicher es ist, dass die Häufigkeiten unterschiedlich sind, als dass sie gleich sind (Manning & Schütze, 1999).

3.5 BERECHNUNG DER LOG-LIKELIHOOD

Die Berechnung der Log-Likelihood erfolgt anhand der erhobenen Häufigkeitsangaben aus beiden Korpora. Für jedes einzelne Wort wird die Log-Likelihood-Ratio berechnet. Zu diesem Zweck wird folgende Kontingenztafel konstruiert:

Tabelle 2: Kontingenztafel zur Berechnung der Log-Likelihood-Ratio

	Korpus 1 = CELEX	Korpus 2 = DeReKo	Gesamt
Häufigkeit Wortform	a	b	a+b
Häufigkeit anderer Wortformen	c-a	d-b	c+d-a-b
Gesamt	c	d	c+d

a...beobachtete Häufigkeit einer bestimmten Wortform in Korpus 1
b...beobachtete Häufigkeit derselben Wortform in Korpus 2
c...Anzahl der Wörter in Korpus 1
d...Anzahl der Wörter in Korpus 2

Daraus wird im nächsten Schritt der Erwartungswert E mit folgender Formel berechnet:

$$E_{\mathrm{i}} = \frac{N_{\mathrm{i}} \sum O_{\mathrm{i}}}{\sum N_{\mathrm{i}}}$$

O……beobachtete Werte (observed values)
N1…..Anzahl der Wörter in Korpus 1 (c)
N2…..Anzahl der Wörter in Korpus 2 (d)

Daraus ergibt sich für jedes einzelne Wort E1=c*(a+b)/(c+d) und E2=d*(a+b)/(c+d). Bei dieser Art der Berechnung der Erwartungswerte wird die Größe des jeweiligen Korpus berücksichtigt (Rayson & Garside, 2000). Unterschiedliche Korpusgrößen haben somit keinen störenden Einfluss auf die Berechnung.

Die Log-Likelihood kann nun für jedes einzelne Wort folgendermaßen berechnet werden: LL = 2*((a*log(a/E1)) + (b*log(b/E2))) (Rayson & Garside, 2000).

Die aus den oben beschriebenen Berechnungen erhaltenen Werte zeigen nun direkt an, wie stark die Häufigkeitsangaben in beiden Korpora bei jedem einzelnen Wort voneinander abweichen. Niedrige Werte bedeuten geringe Abweichung, je höher die Log-Likelihood, desto stärker unterscheiden sich die Häufigkeitsangaben in den Korpora (Dunning, 1993). Bei einer Irrtumswahrscheinlichkeit von α=0.01 liegt das Signifikanzniveau bei 6,63, bei α=0.001 bei 15,13 (Rayson, Berridge, & Francis, 2004). Ist der Log-Likelihood-Wert also höher als 6,63, liegt die Wahrscheinlichkeit, dass der Unterschied zwischen den Worthäufigkeitsangaben lediglich zufallsbedingt ist, bei 1%.

3.6 ERGEBNISSE

Bereits auf den ersten Blick zeigt sich, dass sich die Log-Likelihood-Werte eine relativ große Streuung aufweisen und sich in den vier Gruppen deutlich unterscheiden. Besonders hohe Log-Likelihood-Werte treten tendenziell bei den häufigeren Wörtern auf.

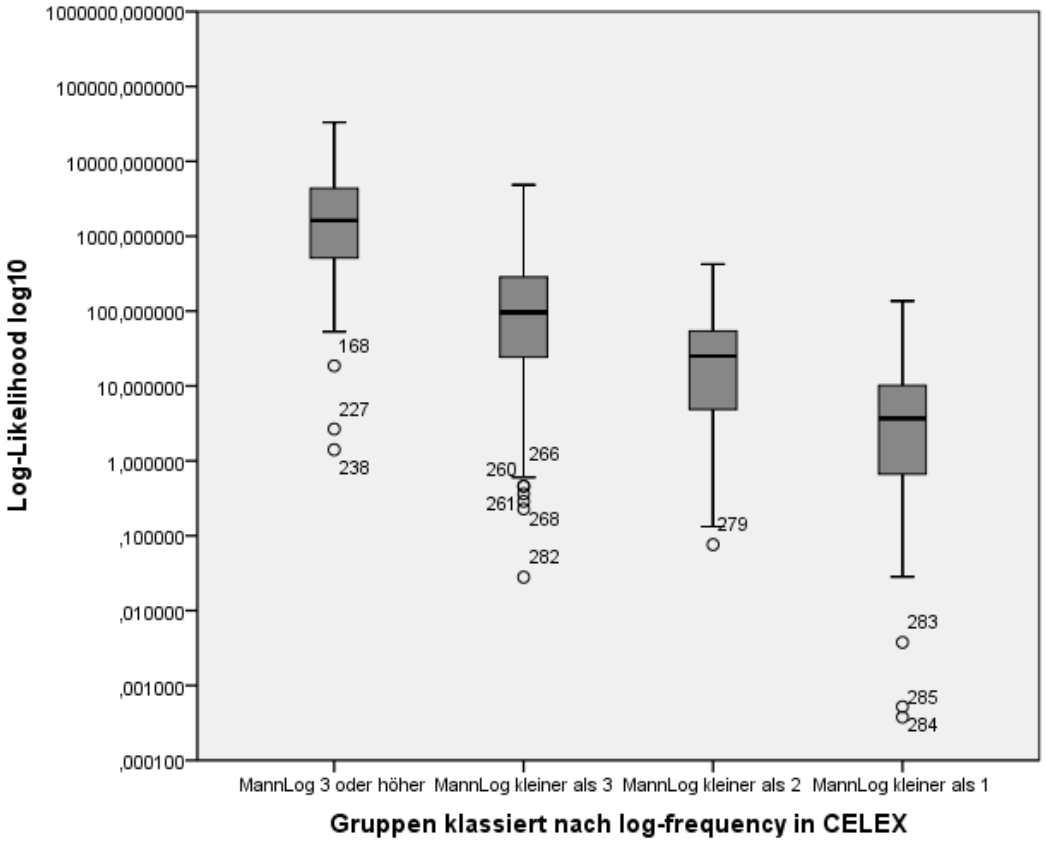

Abbildung 2: Streuung der Log-Likelihood-Werte in den vier Gruppen

Was aus Abbildung 2 bereits anschaulich hervorgeht, zeigt sich auch in den Zahlen in Tabelle 3:

Tabelle 3: Deskriptive Statistiken der Log-Likelihood-Werte in den Gruppen

		Gruppen klassiert nach log-frequency in CELEX			
		Sehr häufig: MannLog 3 oder höher	**Häufig:** MannLog kleiner als 3	**Selten:** MannLog kleiner als 2	**Sehr selten:** MannLog kleiner als 1
Log-Likelihood	Anzahl	60	75	75	75
	Minimum	1,4075	0,0279	0,0760	0,0004
	Maximum	33.211,8098	4.842,0106	423,2248	136,3814
	Median	1.621,1301	96,1997	25,0229	3,6905
	Perzentil 25	524,0747	23,3404	4,7039	0,6191
	Perzentil 75	4.362,9361	287,3879	58,0112	10,3291
	Gültige N	60	75	75	75

Man sieht, dass die Log-Likelihood-Werte in den Gruppen der seltenen und sehr seltenen Wörter eine deutlich geringere Streubreite haben und auch im Mittel – aufgrund der Verteilung der Daten wurde zu diesem Zweck der Median gewählt – deutlich geringere Werte haben. So ist beispielsweise das Maximum in den beiden oben genannten Gruppen um ein Vielfaches niedriger als der Median der Gruppe der sehr häufigen Wörter. Der Median und die Perzentile verdeutlichen die gravierenden Unterschiede zwischen den Gruppen. So ist etwa das Perzentil 25 der Gruppe der sehr häufigen Wörter viel höher als das Perzentil 75 der restlichen Gruppen.

Hohe Log-Likelihood-Werte bedeuten eine hohe Abweichung zwischen den Frequenzangaben in CELEX und DeReKo. Bei einem Signifikanzniveau von $\alpha = 0.01$ gelten Log-Likelihood-Werte größer oder gleich 6,63 als signifikant. Tabelle 4 zeigt die Verteilung signifikanter und nicht-signifikanter Log-Likelihood-Werte in den vier Gruppen:

Tabelle 4: Anteil signifikanter Log-Likelihood-Werte in den Gruppen (absolut und in Prozent)

Gruppe	nicht signifikant	signifikant (α = 0.01)	Gesamt
sehr häufig	2	58	60
	3,30%	96,70%	100,00%
häufig	12	63	75
	16,00%	84,00%	100,00%
selten	22	53	75
	29,30%	70,70%	100,00%
sehr selten	46	29	75
	61,30%	38,70%	100,00%
Gesamt	82	203	285
	28,80%	71,20%	100,00%

Insgesamt erreichen 203 der 285 untersuchten Wörter signifikante Log-Likelihood-Werte, das sind 71,2% der Fälle. Betrachtet man die Gruppen getrennt, zeigt sich, dass die meisten signifikanten Werte bei den häufigeren Wörtern zu finden sind. Während in der Gruppe der sehr seltenen Wörter 38,7% das Signifikanzniveau überschreiten, sind es in der Gruppe der sehr häufigen Wörter 96,7% - lediglich 2 Wörter dieser Gruppe haben nicht-signifikante Log-Likelihood-Werte!

Abbildung 3 stellt die Unterschiede zwischen den vier Gruppen graphisch dar:

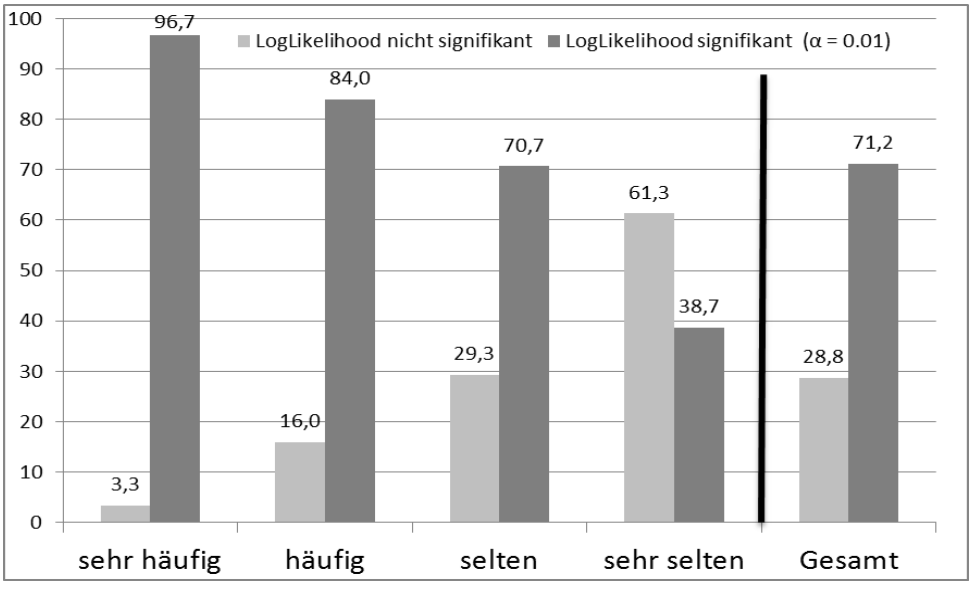

Abbildung 3: Anteil signifikanter Log-Likelihood-Werte in den Gruppen und der Gesamtstichprobe (in Prozent)

Die folgenden Abbildungen zeigen, inwiefern CELEX die Worthäufigkeiten im Vergleich zu den Häufigkeitsangaben im DeReKo über- bzw. unterschätzt.

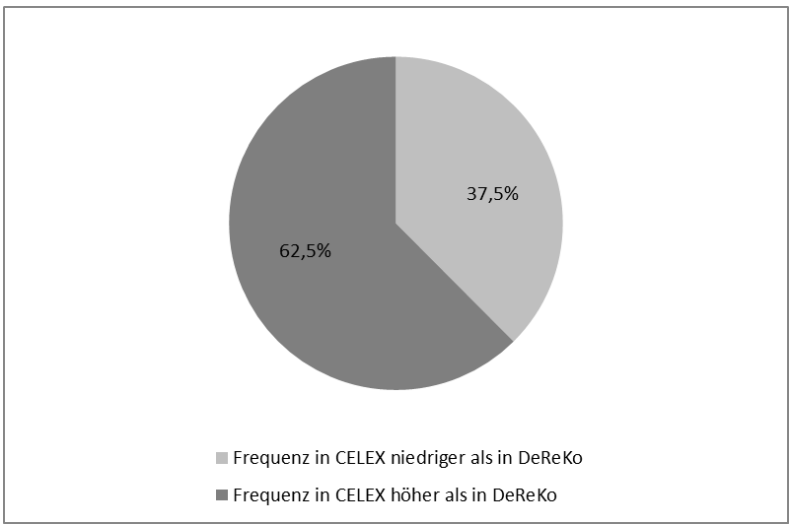

Abbildung 4: Verhältnis zwischen über- und unterschätzten Wortfrequenzen in der Gesamtstichprobe

In 62% der Fälle schätzt CELEX die Worthäufigkeit höher ein als das DeReKo, in 38% der Fälle wird die Worthäufigkeit von CELEX niedriger angegeben, als es im DeReKo der Fall ist.

Abbildung 5 illustriert die Aufteilung zwischen über- und unterschätzten Worthäufigkeiten in den vier Gruppen:

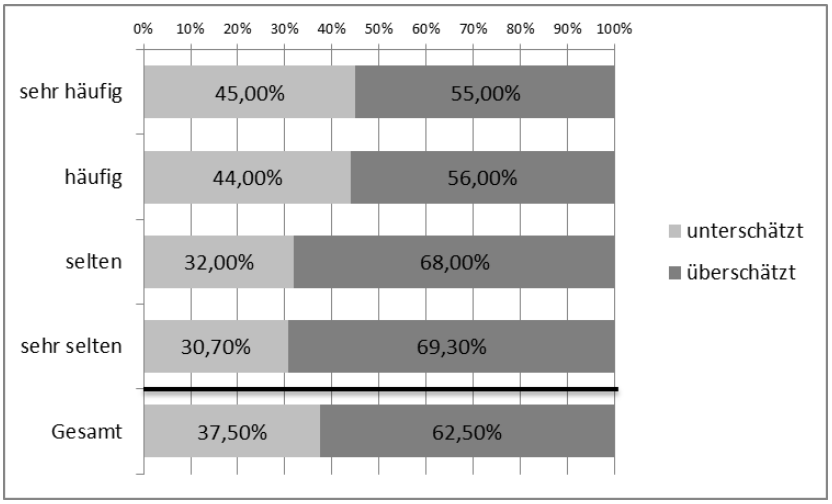

Abbildung 5: Verhältnis zwischen über- und unterschätzten Wortfrequenzen in den Gruppen

Es zeigt sich eine leichte Tendenz zu einer Überschätzung der Wortfrequenz durch CELEX, vor allem in den Gruppen der seltenen und sehr seltenen Wörter.

Betrachtet man nur die Wörter mit einem signifikanten Log-Likelihood-Wert – also jene Wörter, bei denen die Häufigkeitsangaben aus den beiden Datenbanken signifikant voneinander abweichen – ergibt sich folgendes Bild (Abbildung 6):

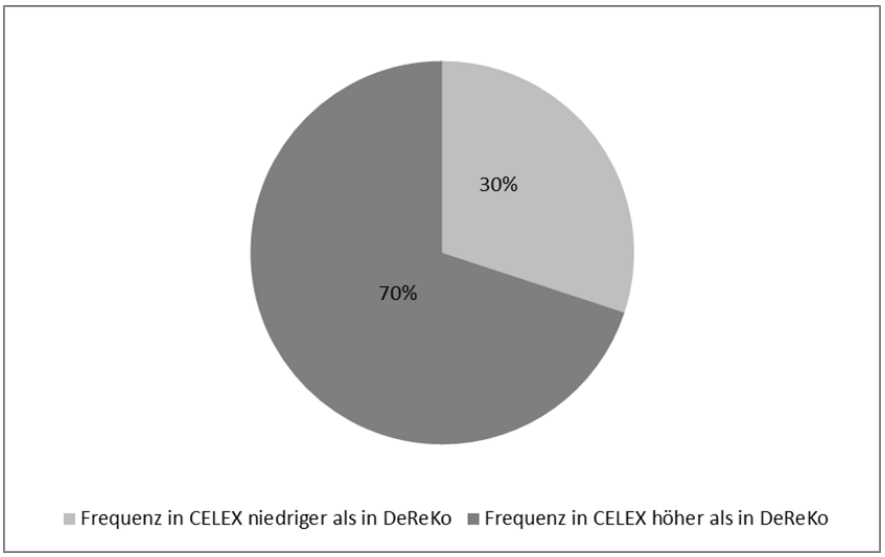

Abbildung 6: Tendenz zur Über- bzw. Unterschätzung der Wortfrequenz bei signifikant abweichenden Häufigkeitsangaben

Bei den signifikant voneinander abweichenden Häufigkeitsangaben schätzt CELEX die Worthäufigkeit in 70% der Fälle höher ein als das DeReKo, bei 30% der Wörter wird die Häufigkeit im Vergleich zum DeReKo unterschätzt.

Abbildung 7 illustriert die Unterschiede zwischen den Gruppen: in den Untergruppen zeigt sich, dass CELEX vor allem bei den seltenen und sehr seltenen Wörtern dazu tendiert, die Worthäufigkeit zu überschätzen. In der Gruppe der seltenen Wörter findet sich mit 84,9% der größte Anteil an überschätzten Frequenzen, in der Gruppe der sehr seltenen Wörter liegt der Anteil bei 75,9%.

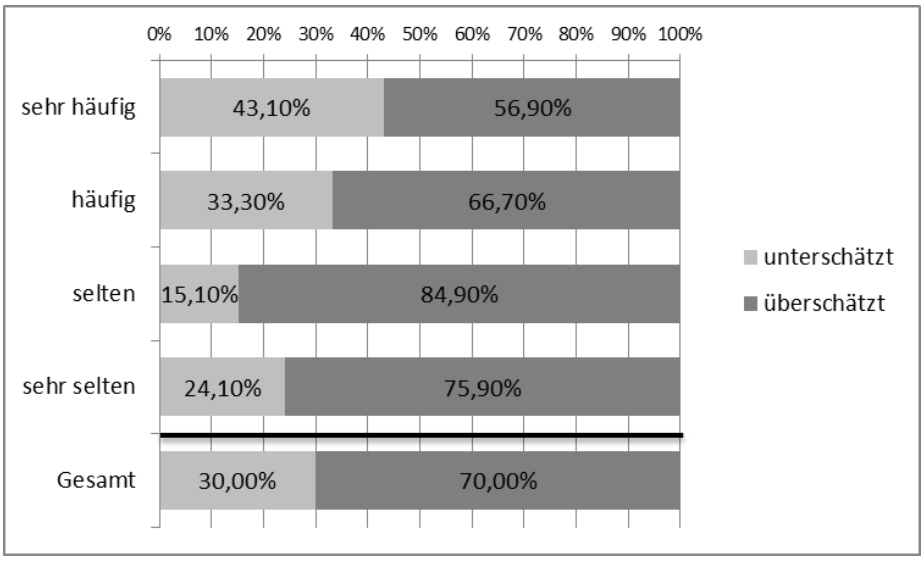

Abbildung 7: Über- bzw. Unterschätzung der Wortfrequenz bei signifikant abweichenden Häufigkeiten - Verteilung in den Gruppen

Ferner ist von Interesse ob es durch die Abweichungen der Häufigkeitsangaben in den beiden Korpora dazu kommt, dass einzelne Wörter beispielsweise im einen Korpus als selten und im anderen als häufig eingestuft werden. Es zeigt sich dass 232 der insgesamt 285 untersuchten Wörter – das entspricht 81,4% - nach den in der vorliegenden Arbeit angewandten Kriterien in beiden Datenbanken derselben Gruppe zugeordnet werden können. Um einen Überblick zu erhalten wie stark es zu Verschiebungen zwischen den Gruppen kommt wurde eine Kreuztabelle erstellt (Tabelle 5):

Tabelle 5: Verschiebung zwischen Gruppen

		DeReKo				
		sehr häufig	häufig	selten	sehr selten	Gesamt
CELEX	sehr häufig	54	6	0	0	60
		90,00%	10,00%	0,00%	0,00%	100,00%
	häufig	1	52	22	0	75
		1,30%	69,30%	29,30%	0,00%	100,00%
	selten	0	1	55	19	75
		0,00%	1,30%	73,30%	25,30%	100,00%
	sehr selten	0	0	4	71	75
		0,00%	0,00%	5,30%	94,70%	100,00%

Tabelle 5 zeigt, dass in den meisten Fällen – zwischen 69,3 Prozent bei den häufigen Wörtern und 94,7 Prozent bei den sehr seltenen Wörtern – die Wörter nach dem in der vorliegenden Arbeit festgelegten Kriterium in beiden Korpora in der gleichen Gruppe eingestuft werden. Sprünge zwischen den Gruppen gibt es nur in die jeweils benachbarte Gruppe. Die größte Verschiebung findet sich bei der Gruppe der in CELEX häufigen Wörter: 69,3 Prozent dieser Wörter werden laut DeReKo ebenfalls als häufig eingestuft, 29,3 Prozent treten dagegen im DeReKo nur selten auf („häufig“ und „selten“ bezieht sich hierbei auf die für die vorliegende Arbeit festgelegte Unterteilung anhand der log-frequency). Bei den in CELEX seltenen Wörtern werden 73,3 Prozent auch laut DeReKo als selten eingestuft, 25,3 Prozent hingegen treten im DeReKo sehr selten auf.

4 DISKUSSION

In den folgenden Abschnitten erfolgt eine Interpretation und Diskussion der Ergebnisse der Log-Likelihood-Analyse, sowie eine kritische Betrachtung der Durchführung der Untersuchung. Abschließend erfolgt ein Ausblick auf die mögliche Verwendung der Erkenntnisse für zukünftige Forschungsarbeiten.

4.1 INTERPRETATION DER ERGEBNISSE

Die Auswertung der Log-Likelihood-Analyse konnte die Hypothese bestätigen, dass die Worthäufigkeitsangaben in CELEX sich von den Häufigkeitsangaben des DeReKo deutlich unterscheiden. Von 285 untersuchten Wörtern konnte bei 71,2% eine signifikante Abweichung festgestellt werden. Dies steht im Einklang mit den Erkenntnissen von Brysbaert und New (2009), welche zeigen konnten, dass sich ein geringer Korpusumfang negativ auf die Qualität von Worthäufigkeitsschätzungen auswirkt. Im Gegensatz zu den Ergebnissen von Burgess und Livesay (1998) und Brysbaert, Buchmeier et al. (2011), welche feststellten, dass bei geringer Korpusgröße vor allem niederfrequente Wörter fehleingeschätzt werden, zeigte sich in der vorliegenden Untersuchung, dass die Frequenzangaben in CELEX vor allem bei den hochfrequenten Wörtern stark von den Angaben im DeReKo abweichen. So zeigten sich in der Gruppe der sehr häufigen Wörter (Vorkommen 1.000 Mal pro Million oder häufiger) besonders starke Abweichungen – bei einer Irrtumswahrscheinlichkeit von 1% war die Log-Likelihood-Ratio bei 96,7% der Wörter signifikant. Auch in der Gruppe der häufigen Wörter (Vorkommen 100 bis 1.000 Mal pro Million) waren noch 84% der Log-Likelihood-Werte signifikant. Bei den sehr seltenen Wörtern (Vorkommen weniger als 10 Mal pro Million) zeigten lediglich 38,7 Prozent eine signifikante Abweichung von den Frequenzangaben aus dem DeReKo.

In Bezug auf die Frage, ob CELEX die Wortfrequenzen im Vergleich zum DeReKo über- oder unterschätzt, zeigte sich im Allgemeinen eine Tendenz zur Überschätzung der Häufigkeiten. In der Gesamtstichprobe wurden die Wortfrequenzen in 62% der Fälle von CELEX überschätzt. Analysiert man die Gruppen getrennt voneinander, zeigt sich eine leichte Tendenz zur Überschätzung der Häufigkeiten durch CELEX, vor allem bei seltenen und sehr seltenen Wörtern. Dieser Effekt verstärkt sich, wenn nur die signifikant abweichenden Frequenzen betrachtet werden. Hier kommt es insgesamt in 70% der Fälle zu einer Überschätzung der Häufigkeit. Getrennt nach Gruppen zeigt sich, dass dieser Effekt bei seltenen und sehr seltenen Wörtern besonders stark ausgeprägt ist: in der Gruppe der seltenen Wörter werden 84,9% der Frequenzen im Vergleich zum DeReKo überschätzt, in der Gruppe der sehr seltenen Wörter immerhin noch 75,9%. Die Ergebnisse hinsichtlich der Richtung der Fehleinschätzung decken sich somit nicht mit der Vermutung von Zevin und Seidenberg (2002), die davon ausgehen dass CELEX aufgrund der Unausgewogenheit des Korpus vor allem die Frequenzen besonders häufiger Wörter unterschätzt. Zu einer Unterschätzung kam es zwar bei den häufigen Wörtern öfter als bei den seltenen Wörtern, allgemein geht die Tendenz jedoch zu einer Überschätzung der Häufigkeit durch CELEX. Die Richtung der Fehleinschätzung von Worthäufigkeiten wurde in den restlichen in der vorliegenden Arbeit zitierten Untersuchungen nicht ausführlich analysiert, es gibt also derzeit weder vergleichbare Daten noch mögliche Erklärungsansätze zu diesem Phänomen.

Ferner zeigte sich, dass es durch die in voneinander abweichenden Häufigkeiten zu Verschiebungen zwischen den für die vorliegende Arbeit eingeteilten Gruppen kommt, allerdings in verhältnismäßig geringem Ausmaß. Mehr als 80% der Wörter bleiben trotz der großen Zahl an signifikanten Log-Likelihood-Werten in der gleichen Häufigkeitsgruppe. Zur größten Verschiebung kommt es bei den nach CELEX als häufig eingestuften Wörtern: 29,3% würden nach der Angabe im DeReKo als selten eingestuft werden.

4.2 KRITIK

Die Aussagekraft der Ergebnisse wird durch den relativ geringen Stichprobenumfang begrenzt, interessanter wäre ein großangelegter Korpusvergleich, welcher jedoch aufgrund der Urheberrechtsproblematik des DeReKo im Fall der vorliegenden Arbeit nicht möglich war.

In der vorliegenden Arbeit wurden für die Stichprobe ausschließlich Wortformen ausgewählt, für zukünftige Studien könnte eventuell die alternative oder zusätzliche Analyse der Frequenzen von Lemmata noch weiteren Informationsgewinn bringen.

Vorangegangene Studien lieferten Hinweise darauf, dass veraltete Quellen eines Korpus nicht die sprachliche Realität der heutigen Zeit widerspiegeln, es ist also anzunehmen, dass sich der Sprachgebrauch und somit auch die tatsächliche Gebrauchshäufigkeit von Wörtern in der Sprache über die Zeit verändern. Aufgrund des geringen Stichprobenumfangs der vorliegenden Arbeit ist eine qualitative Analyse der einzelnen Wörter und deren Frequenzen nicht möglich. Es wäre also eine Anregung für zukünftige Untersuchungen, die Veränderung von Wortfrequenzen inhaltlich zu betrachten, um so zu Erkenntnissen über die Veränderung des Sprachgebrauchs zu gelangen.

Das Deutsche Referenzkorpus DeReKo ist CELEX hinsichtlich des Korpusumfangs, der Aktualität und der Vielfalt der Quellen zweifellos überlegen, jedoch stellt sich die Frage, ob die Frequenzen im DeReKo die tatsächliche Gebrauchshäufigkeit von Wörtern in der alltäglichen Sprache wiederspiegeln. Diese Frage stellt sich allerdings in Bezug auf jedes Korpus.

4.3 AUSBLICK

Wie die Arbeiten von Burgess und Livesay (1998), Brysbaert und New (2009) und Brysbaert, Buchmeier et al. (2011) zeigen konnten, liegt die Zukunft der Ermittlung von Worthäufigkeiten eher im webbasierten Bereich. Vor allem die aus Untertiteln ermittelten sogenannten SUBTLEX-Frequenzen wirken vielversprechend.

Die vorliegende Arbeit schließt an die Ergebnisse oben genannter Untersuchungen an. Die Schlussfolgerung daraus und Empfehlung an Autoren zukünftiger Arbeiten ist, dass es unerlässlich ist, valide und reliable Worthäufigkeitsangaben zu verwenden, und dass aus diesem Grund die Quellen für Worthäufigkeitsangaben mit äußerster Sorgfalt ausgewählt werden müssen.

5 ZUSAMMENFASSUNG

Es gilt als gesichert, dass die Häufigkeit eines Wortes in einer Sprache verschiedene Bereiche der Worterkennung und –verarbeitung beeinflusst. Aus dieser Tatsache ergibt sich die Notwendigkeit zuverlässiger Häufigkeitsangaben. Die vorliegende Arbeit versuchte zu ermitteln, ob die Worthäufigkeitsangaben in den linguistischen Datenbanken CELEX und DeReKo signifikant voneinander abweichen. Da ein großangelegter Vergleich beider Korpora aufgrund urheberrechtlicher Bestimmungen des DeReKo nicht möglich war, wurde eine Zufallsstichprobe von 285 Wörtern aus CELEX gezogen und die korrespondierenden Frequenzangaben aus dem DeReKo ermittelt. Für jedes Wort wurde der Log-Likelihood-Wert errechnet, anhand dessen festgestellt werden konnte, ob eine signifikante Abweichung der Worthäufigkeitsangaben in den beiden Korpora vorliegt. Bei mehr als 70% der untersuchten Wörter konnte eine signifikante Abweichung der Wortfrequenzen nachgewiesen werden. Es zeigte sich, dass es im Bereich der häufigen Wörter zu besonders großen Abweichungen kommt: während in der Gruppe der sehr seltenen Wörter 38,7% einen signifikanten Log-Likelihood-Wert erreichten, waren es in der Gruppe der sehr häufigen Wörter 96,7%. Ferner wurde untersucht, ob CELEX zu einer systematischen Über- oder Unterschätzung der Wortfrequenzen tendiert. Es zeigte sich eine Tendenz zur Überschätzung der Häufigkeiten durch CELEX: in 62% der Fälle wurde die Worthäufigkeit in CELEX höher eingeschätzt als im DeReKo. Bei den Wörtern mit signifikantem Log-Likelihood-Wert wurde die Frequenz in 70% der Fälle von CELEX im Vergleich zum DeReKo überschätzt.

6 ABSTRACT

It is widely acknowledged that the frequency of a word in language influences various areas of word recognition and word processing. Therefore reliable frequency counts are essential. This Thesis tries to explore if the word frequency counts in the linguistic database CELEX differ significantly from the frequency counts in the German Reference Corpus DeReKo. Due to copyright restrictions of the DeReKo a comparison of the entire corpora was not possible, so a randomized sample of 285 wordforms was drawn from CELEX, and the corresponding frequencies were extracted from the DeReKo one by one. For each single word the log-likelihood-ratio was calculated, and used as a measure to judge if there is a statistically significant deviation. More than 70% of the words showed a significant deviation of the frequency counts of the two corpora. Highly frequent words showed even more deviation than low-frequency words: while in the sample of lowest frequency words 38,7% displayed statistically significant log-likelihood-ratios, in the sample of highly frequent words the amount rose to 96,7%. It was further examined if CELEX tends to overestimate or underestimate word frequency. The results showed that CELEX has a tendency to overestimate word frequencies: 62% of the frequencies were overestimated in comparison to the DeReKo. Within the words that had statistically significant log-likelihood-ratios 70% of the frequencies were overestimated by CELEX.

7 LITERATURVERZEICHNIS

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (CD-ROM). Philadelphia, PA.

Balota, D. A., & Chumbley, J. I. (1984). Are Lexical Decisions a Good Measure of Lexical Access? The Role of Word Frequency in the Neglected Decision Stage. *Journal of Experimental Psychology: Human Perception and Performance, 3*(10), pp. 340-357.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology, 2*(133), pp. 283-316.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual Word Recognition: The Journey from Features to Meaning (A Travel Update). In M. J. Traxler, & M. A. Gernsbacher, *Handbook of Psycholinguistics (2nd ed.)* (pp. 285-375).

Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations* (pp. 87-90). Association for Computational Linguistics.

Belica, C., Keibel, H., Kupietz, M., & Perkuhn, R. (2007). Web As Corpus: Kooperation mit der Universität Bologna. *IDS Sprachreport Sonderheft*, pp. 21-25.

Berlin-Brandenburgische Akademie der Wissenschaften. (2012). *DWDS - Digitales Wörterbuch der Deutschen Sprache*. Retrieved 03 05, 2013, from http://www.dwds.de/projekt/hintergrund/

Best, K.-H. (2006). *Quantitative Linguistik.* Göttingen: Peust & Gutschmidt Verlag.

Bibliographisches Institut GmbH. (2013). *Duden.de*. Retrieved 03 12, 2013, from http://www.duden.de/sprachwissen/sprachratgeber/die-haeufigsten-woerter-in-deutschsprachigen-texten

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using internet search engines to estimate word frequency. *Behaviour Research Methods, Instruments and Computers, 2*(34), pp. 286-290.

Bopp, S. (2010). *Einführung in die Korpusliguistik mit DeReKo und COSMAS II.* Retrieved 03 05, 2013, from http://www.philhist.uni-

augsburg.de/de/lehrstuehle/germanistik/sprachwissenschaft/mitarbeiter/stelspass/materialien_lehrveranstaltungen/korpuslinguistik_dereko_cosmas2_bopp.pdf

Brybaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect - A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology, 5*(58), pp. 412-424.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical Evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behaviour Research Methods, 4*(41), pp. 927-990.

Brysbaert, M., Lange, M., & Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurence in visual word recognition: further evidence from the Dutch language. *European Journal of Cognitive Psychology, 12*, pp. 65-85.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recogition task: Moving on from Kucera and Francis. *Behaviour Research Methods, Instruments and Computers, 2*(30), pp. 272-277.

Burnage, G. (1990). *CELEX - A Guide for Users.* Nijmegen: CELEX Centre For Lexical Information.

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft.* Stuttgart: Alfred KrönerVerlag.

Conrad, M., & Jacobs, A. M. (2004). Replicating syllable frequency effects in Spanish in German: One more challenge to computational models of visual word recognition. *Language and Cognitive Processes, 3*(19), pp. 369-390.

Cortese, M. J., & Balota, D. A. (2012). Visual Word Recognition in Skilled Adult Readers. In M. J. Spivey, K. McRae, & M. F. Joanisse, *The Cambridge Handbook of Psycholinguistics* (pp. 159-185). Cambridge University Press.

Dunning, T. (1993, 03). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, 19*(1), pp. 61-74.

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). WordGen: A Tool for Word Selection and Non-Word Generation in Dutch, English, German and

French. *Behaviour Research, Methods, Instruments and Computers, 36*(3), pp. 488-499.

Forster, K. I. (1976). Accessing the Mental Lexicon. In R. J. Wales, & E. Walker, *New approaches to language mechanisms* (pp. 257-287). North-Holland Publishing Company.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behaviour, 12*, pp. 627-635.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of Acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica, 115*, pp. 43-67.

Graf, R., Nagler, M., & Jacobs, A. M. (2005). Faktorenanalyse von 57 Variablen der visuellen Worterkennung. *Zeitschrift für Psychologie, 4*(213), pp. 205-218.

Grainger, J. (1990). Word Frequency and Neighborhood Frequency Effects in Lexical Decision and Naming. *Journal of Memory and Language, 29*, pp. 228-244.

Gulikers, L., Rattink, G., & Piepenbrock, R. (1995). German Linguistic Guide. In R. Baayen, R. Piepenbrock, & L. Gulikers, *The CELEX Lexical Database (CD-ROM).* Philadelpia, PA: Linguistic Data Consortium.

Hofmann, M. J., Stenneken, P., Conrad, M., & Jacobs, A. M. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behaviour Research Methods, 3*(39), pp. 620-629.

Howes, D. H., & Solomon, R. L. (1951). Visual Duration Threshold as a Function of Word-Probability. *Journal of Experimental Psychology, 6*(41), pp. 401-410.

Institut für Deutsche Sprache. (1991-2010). *COSMAS I/II Corpus Search, Management and Analysis System*. (Mannheim: Institut für Deutsche Sprache) Retrieved 12 08, 2012, from http://www.ids-mannheim.de/cosmas2/

Institut für Deutsche Sprache. (2004). *Grundstrukturen: Freiburger Korpus*. (Mannheim: Institut für Deutsche Sprache) Retrieved 04 04, 2013, from http://dsav-oeff.ids-mannheim.de/DSAv/KORPORA/FR/FR_DOKU.HTM

Institut für Deutsche Sprache. (2012b). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2012-II*. (Mannheim: Institut für Deutsche Sprache) Retrieved 11 11, 2012, from http://www.ids-mannheim.de/DeReKo

Kaeding, F. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache.* Berlin: F. W. Kaeding.

Keibel, H. (2008). *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung.* Retrieved 13 03, 2013, from http://www.ids-mannheim.de/kl/dokumente/freqMeasures.html

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogenity and similarity between corpora. *Proceedings of ACL-SIGDAT Workshop on very large corpora* , 231-245. Beijing and Hong Kong.

Kronbichler, M., Hutzler, F., Wimmer, H., Mair, A., Staffen, W., & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *NeuroImage, 21*, pp. 946-953.

Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi, & Y. Kawaguchi (Eds.), *Working Papers in Corpus-based Linguistics and Language Education* (Vol. 3). Tokyo: Tokyo University of Foreign Studies (TUFS).

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A Primodial Sample for Linguistic Research. *Proceedings of the 7th conference on international language resources and evaluation (LREC 2010)*, (pp. 1848-1854).

Kupietz, M., Schonefeld, O., & Witt, A. (2010, 05 23). The German Reference Corpus: New developments building on almost 50 years of experience. *Language Resources: From Storyboard to Sustaiability and LR Lifecycle Management*, *39.*

Lemnitzer, L., & Zinsmeister, H. (2010). *Korpuslinguistik - Eine Einführung.* Tübingen: Narr Francke Attempto Verlag.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurences. *Behaviour Research Methods, Instruments and Computers, 2*(28), pp. 203-208.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Meier, H. (1964). *Deutsche Sprachstatistik.* Hildesheim: Georg Olms Verlagsbuchhandlung.

Morrison, C. M., & Ellis, A. W. (1995). Roles of Word Frequency and Age of Acquisition in Word Naming and Lexical Decision. *Journal of Experimental Psychology: Learning, Memory and Cognition, 1*(21), pp. 116-133.

Murray, W. S., & Forster, K. I. (2004). Serial Mechanisms in Lexical Access: The Rank Hypothesis. *Psychological Review, 3*(111), pp. 721-756.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*, pp. 661-677.

Quasthoff, U. (2009). Korpusbasierte Wörterbucharbeit mit den Daten des Projekts Deutscher Wortschatz. *Linguistik online, 39*(3).

Rayson, P., & Garside, R. (2000). Comparing Corpora using Frequency Profiling. *Proceedings of the workshop on Comparing Corpora*, (pp. 1-6). Association for Computational Linguistics.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. *7th International Conference on Statistical analysis of textual data (JADT 2004)*, (pp. 926-936).

Spieler, D. H., & Balota, D. A. (2000). Factors Influencing Word Naming in Younger and Older Adults. *Psychology and Aging, 15*(2), pp. 225-231.

Universität Regensburg. (2006). *DWDS Kurzanleitung.* Retrieved 03 05, 2013, from http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Korpuslinguistik/materialien/dwds.pdf

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behaviour, 17*(2), pp. 143-154.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47*, pp. 1-29.

8 ANHANG

8.1 ABBILDUNGS- UND TABELLENVERZEICHNIS

ıch habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit einzuholen. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

8.2 WORTFORMEN UND HÄUFIGKEITEN

8.2.1 GRUPPE 1: SEHR HÄUFIGE WÖRTER

Logarithmierte Häufigkeit pro Million (log-frequency) in CELEX 3 oder höher

Wortform	CELEX frequency per million	CELEX log-frequency	DeReKo frequency per million	DeReKo log-frequency	Log-Likelihood
und	25287	4,402897	23076,71	4,363174	4209,5430
in	18536	4,268016	18199,37	4,260056	1657,3053
der	12255	4,088313	34023,02	4,531773	30748,7796
die	11474	4,059715	33603,16	4,52638	33211,8098
nicht	8022	3,904283	5647,55	3,75186	4818,1585
ist	7453	3,872331	6896,93	3,838656	1119,0671
es	6428	3,808076	5453,54	3,736678	1698,9357
dass	5912	3,771734	2283,87	3,358671	13701,4237
ich	5857	3,767675	1881,54	3,274513	17483,7656
er	5803	3,763653	3772,27	3,576603	4516,3292
zu	5180	3,71433	7883,55	3,896722	881,3207
auch	4845	3,685294	5544,9	3,743894	18,5953
von	4693	3,671451	9246,27	3,965966	3670,2279
mit	4304	3,633872	9729,86	3,988107	5734,8612
den	4031	3,605413	11541,05	4,062245	10991,5320
das	3832	3,583426	10268,94	4,011525	8709,7760
auf	3705	3,568788	7346,32	3,86607	2979,3806
werden	3598	3,556061	3583,9	3,554355	276,5307
sie	3507	3,544936	4102,02	3,612998	2,6600
wir	3460	3,539076	2122,47	3,326842	3176,8896
so	3380	3,528917	2619,44	3,418209	1395,2992
hat	3356	3,525822	3185,26	3,503146	410,9497
fuer	3308	3,519566	8044,5	3,905499	5619,1943
war	3261	3,513351	2562,97	3,408744	1265,7051
noch	3199	3,505014	2785,24	3,444862	727,3750
wird	3187	3,503382	3109,23	3,492652	303,5599
bei	2994	3,476252	4142,25	3,617236	185,0966
sind	2961	3,471438	2859,41	3,456277	311,1398
nur	2899	3,462248	2236,16	3,349503	1220,7900
des	2790	3,445604	6470,1	3,810911	4058,4249
man	2738	3,437433	1669,69	3,222637	2554,5313
haben	2517	3,400883	2151,86	3,332814	637,1996

an	2504	3,398634	4641,51	3,666659	1481,0136
als	2423	3,384353	4426,24	3,646035	1333,2020
sich	2387	3,377852	6885,24	3,837919	6637,6429
wenn	2364	3,373647	1296,47	3,112763	2869,9098
ja	2352	3,371437	342,21	2,534293	14806,4610
oder	2225	3,34733	1853,16	3,267913	650,1522
dem	2195	3,341435	6282,16	3,798109	5980,1038
aus	1940	3,287802	4239,72	3,627337	2289,6772
wie	1802	3,255755	2756,6	3,440374	320,2115
dann	1802	3,255755	1157	3,063334	1453,2237
diese	1797	3,254548	1036,25	3,015465	1935,5395
hatte	1794	3,253822	1355,55	3,132116	824,1952
aber	1745	3,241795	2510,18	3,399704	173,0125
nach	1738	3,24005	3982,73	3,60018	2426,8997
wurde	1688	3,227372	1972,53	3,295023	1,4075
kann	1684	3,226342	1227,04	3,088857	892,6668
ein	1550	3,190332	6893,06	3,838412	11354,7177
dieser	1540	3,187521	887,82	2,948323	1659,4429
schon	1458	3,163758	1309,32	3,117046	271,6691
eine	1449	3,161068	5586,35	3,747128	7964,7070
wieder	1262	3,101059	1303,78	3,115206	61,8118
jetzt	1253	3,097951	804,73	2,905652	1010,6543
vor	1244	3,09482	2776,28	3,443463	1584,9548
hier	1218	3,085647	588,48	2,769731	1931,8516
durch	1199	3,078819	1621,59	3,209942	53,1854
immer	1162	3,065206	940,2	2,973218	393,1097
koennen	1144	3,058426	1102,38	3,042331	122,2643
also	1089	3,037028	291,59	2,464772	4021,6339

8.2.2 GRUPPE 2: HÄUFIGE WÖRTER

Logarithmierte Häufigkeit pro Million in CELEX kleiner als 3

Wortform	CELEX frequency per million	CELEX log-frequency	DeReKo frequency per million	DeReKo log-frequency	Log-Likelihood
sehr	976	2,98945	694,94	2,841946	564,0511
habe	954	2,979548	950,07	2,977755	73,3094
waren	846	2,92737	902,3	2,955349	25,1481
da	768	2,885361	929,98	2,968473	0,6023
ganz	763	2,882525	603,4	2,780609	287,3879
du	694	2,841359	130,21	2,114654	3590,6290
neue	628	2,79796	816,73	2,91208	13,1274
wuerde	602	2,779596	280,59	2,448077	1019,9774
bis	576	2,760422	2755,49	3,440199	4842,0106
vielleicht	466	2,668386	187,94	2,274025	1009,4036
heute	443	2,646404	853,81	2,931362	314,7564
sollte	414	2,617	359,3	2,555455	96,1997
kommt	406	2,608526	425,91	2,629317	16,0971
zwar	405	2,607455	400	2,602062	34,4785
darauf	400	2,60206	256,85	2,409676	322,9025
werde	365	2,562293	397,51	2,599352	7,2750
mal	356	2,55145	536,61	2,72966	56,0979
davon	351	2,545307	305,02	2,484326	80,9532
dafuer	346	2,539076	419,96	2,623207	0,3633
Bundesrepublik	346	2,539076	32,94	1,517744	2851,4235
Hand	316	2,499687	153,55	2,186245	493,7844
Beispiel	315	2,498311	217,5	2,33745	202,3529
Sowjetunion	298	2,474216	13,11	1,117585	3564,9558
seit	294	2,468347	979,02	2,990792	1169,7581
sogar	287	2,457882	312,15	2,494363	5,9353
seinem	279	2,445604	549,28	2,739791	218,6815
weiter	258	2,41162	481,17	2,682299	157,7746
politischen	257	2,409933	98,69	1,994251	598,4136
lassen	242	2,383815	419,98	2,623232	101,7789
zweiten	238	2,376577	365,5	2,562891	43,9800
bitte	234	2,369216	60,61	1,782517	896,5923
Mittwoch	211	2,324282	309,07	2,490053	25,7729
denen	210	2,322219	251,48	2,400496	0,0279
seiner	191	2,281033	809,71	2,908328	1272,0711
meine	187	2,271842	165,06	2,21764	39,5449

richtig	187	2,271842	152,87	2,184311	59,8634
Stunden	181	2,257679	204,04	2,309718	1,2757
gekommen	181	2,257679	147,99	2,170232	58,7579
kleine	176	2,245513	202,32	2,306037	0,4589
liegen	174	2,240549	154,04	2,187642	35,9780
Recht	169	2,227887	196,74	2,293889	0,2278
Mark	168	2,225309	352,88	2,547631	171,3108
Gespraech	165	2,217484	116,82	2,067508	97,5851
amerikanische	165	2,217484	35,73	1,553068	751,0183
obwohl	163	2,212188	159,68	2,203238	14,7137
Verhandlungen	162	2,209515	63,76	1,804574	365,7037
hielt	158	2,198657	82,51	1,91652	214,1797
Raum	157	2,1959	122,42	2,087858	64,0504
Industrie	157	2,1959	54,89	1,739454	418,1175
grosser	147	2,167317	99,89	1,999536	99,5571
Einheit	146	2,164353	30,84	1,489089	682,8208
Spiel	143	2,155336	364,38	2,561559	279,3232
gefunden	140	2,146128	127,69	2,106153	23,3404
haelt	137	2,136721	168,01	2,225343	0,2869
Macht	135	2,130334	342,09	2,534145	261,4595
Technik	134	2,127105	63,08	1,799877	224,2509
neuer	132	2,120574	150,25	2,176827	0,6708
gemeinsam	131	2,117271	211,32	2,324936	35,2042
erklaerte	126	2,100371	225,22	2,352609	63,0264
mag	126	2,100371	56,82	1,754534	225,1889
zunaechst	124	2,093422	202,52	2,30646	34,9723
Sinne	121	2,082785	52,28	1,718349	232,6982
Preis	120	2,079181	130,21	2,114634	2,6936
Liebe	114	2,056905	115,55	2,062776	7,5150
Tuer	113	2,053078	76,87	1,885746	75,9729
uebrigen	113	2,053078	55,06	1,740865	177,9868
handelt	111	2,045323	82,45	1,916197	55,5932
Tatsache	110	2,041393	45,56	1,65856	228,0999
versucht	108	2,033424	82,51	1,916498	48,2636
Vorsitzenden	107	2,029384	56,88	1,754988	137,2934
britischen	103	2,012837	35,47	1,549868	281,9804
bevor	102	2,0086	110,6	2,04374	2,1935
bringt	101	2,004321	114,77	2,059834	0,4601
Angst	100	2	89,89	1,953708	18,7061
bisschen	100	2	43,38	1,637306	195,0353

8.2.3 GRUPPE 3: SELTENE WÖRTER

Logarithmierte Häufigkeit pro Million in CELEX kleiner als 2

Wortform	CELEX frequency per million	CELEX log-frequency	DeReKo frequency per million	DeReKo log-frequency	Log-Likelihood
wissenschaftlichen	73	1,863323	12,43	1,09434	404,4484
folgende	71	1,851258	43,73	1,640794	64,7133
hoehere	64	1,80618	45,49	1,657918	36,7514
Dingen	62	1,792392	14,97	1,175247	253,3692
technische	59	1,770852	37,66	1,575869	47,2894
Lager	56	1,748188	32,9	1,517238	58,0112
Gewerkschaft	54	1,732394	30,66	1,486561	62,1120
Umstaenden	54	1,732394	19,76	1,295808	137,0800
meiner	52	1,716003	75,46	1,877725	5,7466
Journalisten	50	1,69897	33,42	1,524051	36,3931
befindet	44	1,643453	55,89	1,747335	0,4900
Kraeften	44	1,643453	12,34	1,091469	151,9390
Rennen	37	1,568202	98,94	1,995375	84,4571
still	37	1,568202	16,37	1,21414	68,6707
Gebrauch	37	1,568202	14,93	1,173917	80,9726
Beispiele	35	1,544068	22,42	1,35069	29,3418
geistige	33	1,518514	7,11	0,851799	149,1692
geeignet	30	1,477121	24,44	1,388078	9,9714
getreten	30	1,477121	13,75	1,138218	50,3823
werktaetigen	30	1,477121	0,85	-0,068142	423,2248
Pfarrer	28	1,447158	87,74	1,943206	97,5700
Vaters	28	1,447158	17,7	1,247903	22,7565
Geschaefte	27	1,431364	41,8	1,62117	5,9184
jedesmal	27	1,431364	4,88	0,688697	146,0929
Pferde	26	1,414973	19,39	1,287668	11,4957
zufaellig	26	1,414973	15,92	1,201887	22,4414
erfasst	26	1,414973	14,94	1,174383	29,6917
Taeter	25	1,39794	107,58	2,031747	171,6044
Grab	25	1,39794	11,21	1,049451	46,5130
Gas	24	1,380211	24,18	1,383397	1,2382
beseitigen	24	1,380211	10,23	1,009935	49,5248
Geschehen	23	1,361728	25,67	1,409389	0,1912
entschlossen	23	1,361728	20,3	1,307468	4,0817
endgueltige	23	1,361728	14,08	1,14866	21,8713
ausfuehrliche	23	1,361728	5,63	0,750508	89,2988

Brand	21	1,322219	56,33	1,75074	48,5693
vollem	21	1,322219	12,07	1,081631	21,5836
Eisen	21	1,322219	10,54	1,02276	30,8770
Bundespraesidenten	21	1,322219	8,87	0,947712	41,4405
dahinter	19	1,278754	23,49	1,370892	0,0760
fuerchten	19	1,278754	19,7	1,294391	1,1317
erzeugen	19	1,278754	10,04	1,001546	25,0229
Belgrad	18	1,255273	15,85	1,200158	4,5869
stattgefunden	18	1,255273	12,19	1,086001	13,2484
Unsinn	18	1,255273	6,71	0,826548	44,8647
Disziplin	17	1,230449	18,75	1,273016	0,1739
ueberaus	17	1,230449	18,71	1,272118	0,1323
empfehlen	17	1,230449	14,61	1,16478	4,9954
Heft	17	1,230449	8,18	0,912903	29,3726
Teilung	17	1,230449	4,92	0,692093	61,0024
verantwortlichen	16	1,20412	47,9	1,680367	49,2105
Stall	16	1,20412	11,79	1,071345	8,4600
zugute	15	1,176091	28,75	1,45863	10,4979
treu	15	1,176091	15,05	1,177455	0,8842
Hauptmann	15	1,176091	6,7	0,826356	25,5276
abendlaendischen	15	1,176091	0,79	-0,101331	167,9302
Szenen	14	1,146128	23,81	1,37668	4,7039
gefaehrlichen	14	1,146128	14,85	1,171659	0,7631
beruflichen	13	1,113943	20,68	1,315577	2,7314
einstellen	13	1,113943	18,96	1,277867	1,9422
Gerichte	13	1,113943	12,18	1,08559	2,1242
demselben	13	1,113943	4,76	0,677194	30,2326
schnelle	12	1,079181	25,97	1,414422	12,8602
ergaenzen	12	1,079181	10,92	1,038231	2,4199
strenge	12	1,079181	9,94	0,997306	4,1240
Gesichtspunkten	12	1,079181	3,68	0,565542	36,3131
Stichwort	11	1,041393	19,63	1,292846	4,5129
Melodie	11	1,041393	6,44	0,809195	12,3480
musikalische	10	1	52,46	1,719861	96,2784
voruebergehend	10	1	16,66	1,221723	3,7451
angezogen	10	1	5,62	0,750077	10,4783
Durst	10	1	5,13	0,710325	12,2464
Fluesse	10	1	4,99	0,69845	16,7528
halber	10	1	4,74	0,676173	15,2127
weigerte	10	1	3,79	0,578744	25,3160

8.2.4 GRUPPE 4: SEHR SELTENE WÖRTER

Logarithmierte Häufigkeit pro Million in CELEX zwischen 1 (jedoch größer als 0)

Wortform	CELEX frequency per million	CELEX log-frequency	DeReKo frequency per million	DeReKo log-frequency	Log-Likelihood
doppelt	9	0,954243	27,36	1,437162	29,9690
demonstriert	9	0,954243	6,46	0,810047	6,1655
Naturwissenschaften	9	0,954243	3,96	0,597353	17,3142
Gehoer	8	0,90309	10,82	1,034396	0,7112
angehoert	8	0,90309	9,24	0,965464	0,0004
gemein	8	0,90309	4,61	0,664049	7,6298
aufheben	8	0,90309	3,46	0,539367	15,8064
bejaht	8	0,90309	1,07	0,027822	52,3863
faschistische	8	0,90309	0,54	-0,266141	79,8848
entspringt	7	0,845098	1,73	0,23871	29,8892
Apfel	6	0,778151	5,94	0,773921	0,3839
gedient	6	0,778151	3,76	0,575749	6,5053
Oberstleutnant	6	0,778151	3,72	0,570437	6,7095
Regionalliga	5	0,69897	24,49	1,38893	47,4198
vorrangig	5	0,69897	6,74	0,828376	0,6191
klugen	5	0,69897	2,19	0,339631	8,0232
Artillerie	5	0,69897	2,08	0,31702	11,5167
radioaktiven	5	0,69897	1,92	0,283936	12,9013
Meteor	5	0,69897	0,35	-0,458027	46,8157
Voraussagen	5	0,69897	0	-2,526213	136,3814
verletzten	4	0,60206	29,32	1,467108	71,9362
Wanderungen	4	0,60206	9,89	0,995138	7,0330
verstaendigt	4	0,60206	8,17	0,91227	4,0319
Vergewaltigung	4	0,60206	7,73	0,888029	2,0327
entzuendet	4	0,60206	4,72	0,673542	0,0475
Fluch	4	0,60206	4,56	0,6594	0,0005
Gewissens	4	0,60206	1,84	0,264688	7,9594
Irre	4	0,60206	1,81	0,256708	5,7305
zuverlaessiger	4	0,60206	1,56	0,192704	10,3291
Raumschiffes	4	0,60206	0,13	-0,870594	53,5791
luden	3	0,477121	8,37	0,922629	9,9458
Wahrzeichen	3	0,477121	6,36	0,803643	3,6141
Paerchen	3	0,477121	4,33	0,636202	0,9925
hergerichtet	3	0,477121	4,03	0,605446	0,1047
treffend	3	0,477121	3,68	0,565938	0,1435

Schwerpunkten	3	0,477121	3,01	0,478162	0,2368
Hoerner	3	0,477121	2,64	0,422384	0,4267
Umwege	3	0,477121	2,61	0,417034	0,2659
schliesse	3	0,477121	2,52	0,400965	0,3676
endlose	3	0,477121	2,3	0,361967	1,7610
streicheln	3	0,477121	2,24	0,350222	0,5031
Extrem	3	0,477121	2,08	0,318342	2,4614
Zoos	3	0,477121	2,08	0,317487	1,1243
gotischen	3	0,477121	1,65	0,218764	3,6905
vorwirft	3	0,477121	1,59	0,20082	4,0443
entsprachen	3	0,477121	1,48	0,169925	5,4744
erhobene	3	0,477121	1,45	0,160759	6,5452
Annehmlichkeiten	3	0,477121	0,92	-0,034676	8,4609
Wirtschaftsjahr	3	0,477121	0,82	-0,084321	13,6040
bebaute	3	0,477121	0,47	-0,328588	18,1689
Abendbrot	3	0,477121	0,47	-0,330313	20,0079
Staatsform	3	0,477121	0,46	-0,333436	16,5756
Podest	2	0,30103	7,65	0,883742	11,7620
mitzubringen	2	0,30103	6,52	0,814331	7,4777
verrechnet	2	0,30103	3,54	0,548786	0,8984
Trachten	2	0,30103	2,81	0,448241	0,3089
gewohnter	2	0,30103	2,45	0,388725	0,4107
Sozialversicherungen	2	0,30103	2,13	0,329079	0,1201
zaghaft	2	0,30103	1,96	0,29127	0,2030
abgetan	2	0,30103	1,74	0,2415	0,0038
zugetragen	2	0,30103	1,65	0,218178	1,4229
bang	2	0,30103	1,65	0,217493	0,0284
kaeuflich	2	0,30103	1,58	0,197961	0,4722
ermuntern	2	0,30103	1,52	0,180612	0,9459
handhaben	2	0,30103	1,39	0,143337	1,8053
grell	2	0,30103	1,24	0,092097	1,3165
hoffnungsvollen	2	0,30103	1,18	0,070659	2,7608
verstoert	2	0,30103	1,16	0,064295	1,0805
eingegeben	2	0,30103	1,12	0,048542	0,7573
Siedlungsgebiet	2	0,30103	1,05	0,020176	0,9601
saugen	2	0,30103	1,05	0,019713	4,3373
verhoehnt	2	0,30103	0,96	-0,017515	1,2594
Strichen	2	0,30103	0,69	-0,160491	2,6581
elektrischem	2	0,30103	0,38	-0,421555	8,8592
Magnetismus	2	0,30103	0,23	-0,646114	9,3413

CURRICULUM VITAE

Persönliche Angaben

Karin Gmoser

geboren am 04. Dezember 1979 in Wien

Ausbildung

1986-1990	Volksschule Alterlaa, 1230 Wien
1990-1994	GRG 23, 1230 Wien
1994-1999	HBLA Bergheidengasse, 1130 Wien
ab 1999	Diplomstudium Psychologie Universität Wien

Berufliche Tätigkeit

2000-2008	TeamTraining Austria
2007-2008	Praktikum bei 147 Rat auf Draht
seit 2008	147 Rat auf Draht

Fremdsprachen

Englisch, Spanisch