Erschienen in: Kratochvílová, Iva/Wolf, Norbert Richard (Hrsg.):
Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer
Perspektive. Heidelberg: Winter, 2010. S. 117-134. (Germanistische Bibliothek 38)

Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie

Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus[digital]

Sabine Krome, Gütersloh

Korpusanalytische Verfahren als Grundlage von Analysen der deutschen Gegenwartssprache sind spätestens seit dem COBUILD-Projekt von 1987 eine vieldiskutierte Methode fundierter lexikographischer Arbeit.[1] Nicht zuletzt aufgrund der Erfahrungen an diesem Vorreiter-Projekt wird heute der ‚korpusgestützten' gegenüber der ‚korpusgebundenen' Lexikographie der Vorzug gegeben, also die Orientierung an lediglich einem Korpus als selbstreferentiellem System weitgehend kritisch betrachtet. Bei der korpusgestützten Vorgehensweise wird ferner zwischen einer ‚korpusgesteuerten' (‚corpus-driven') — „vom Korpus selbst ausgehen[den]" — und einer ‚korpusvalidierenden' oder ‚korpusbasierten' (‚corpus-based')— „das Korpus zur Rückprüfung benutzen[den]" — Methode (Klosa 2007, 113) unterschieden, wobei auch eine Kombination beider Methoden in der Lexikographie Anwendung findet. Darüber hinaus hat sich die Auffassung durchgesetzt, dass rein quantitativ ausgerichtete Analysen mit dem Schwerpunkt auf statistisch orientierten Methoden für eine grundlegende lexikographische Arbeit deutlich zu kurz greifen und durch qualitative Methoden verschiedenster Art ergänzt bzw. differenziert werden müssen.[2] Dies gilt gleichermaßen für die linguistische Analyse wissenschaftlicher Texte mit Spezialkorpora wie für die Erarbeitung allgemeinsprachlicher Wörterbücher. In der Kombination einer differenzierten quantitativen und qualitativen Methodik unter Einbeziehung sekundärer und tertiärer Quellen aber ist die Korpusanalyse aus der modernen Lexikographie nicht mehr wegzudenken.

Bei der Erarbeitung allgemeinsprachlicher Wörterbücher, die das Ziel haben, die Bedeutung von Lexemen und ihre Verwendung im Sprach- und Schreibgebrauch darzustellen, ist der Lexikograph in besonderem Maße darauf angewiesen, die Konventionen einer breiten Mehrheit der Sprachteilnehmer ‚authentisch', also möglichst empirisch nachvollziehbar, abzubilden. In der lexikographischen Praxis, die hier im Vordergrund der Betrachtung stehen soll, spielen dabei drei zentrale Aspekte eine Rolle:

1. die Erweiterung der Stichwortsubstanz durch Neologismen,
2. die Möglichkeit der Aktualisierung der vorhandenen Substanz sowie —

---

[1]Die Entstehung und Entwicklung sowie die korpusanalytischen Grundlagen des COBUILD-Wörterbuchs beschreibt John Sinclair (1987, 1991), dazu auch Engelberg/Lemnitzer (2001, 206).
[2]ur Diskussion der verschiedenen korpusanalytischen Ansätze und Methoden in der Lexikographie vgl. z. B. Klosa (2007), Lüdeling (2007), Mair (1991), Lemnitzer/Zinsmeister (2006, 36/37).

als neuer Untersuchungsgegenstand vor allem im Hinblick auf orthografische Wörterbücher —,

3. die Schreibbeobachtung vor dem Hintergrund der Arbeit des Rats für deutsche Rechtschreibung.

Anhand dieser drei Aufgabenstellungen soll gezeigt werden, welche neuen Perspektiven die Korpusanalyse trotz aller immer noch vorhandenen Einschränkungen und Probleme für aktuelle allgemeinsprachliche Wörterbücher des Deutschen bietet. Hier werden ganz eigene Anforderungen an ein Basiskorpus gestellt. Dies ist sowohl bedingt durch die Zielgruppe wie auch durch die Beschränkungen des Umfangs eines einbändigen Wörterbuchs als Printprodukt — im Gegensatz etwa zu korpusbasiert erarbeiteten elektronischen Wörterbüchern.

1. Allgemeinsprachliche Wörterbücher als Spiegel der deutschen Gegenwartssprache

Allgemeinsprachliche Wörterbücher haben den Anspruch, die deutsche Gegenwartssprache authentisch widerzuspiegeln. Neben einer ‚relativen Vollständigkeit' muss also die Aktualität des Wortschatzes gewährleistet sein, und es müssen Ausgewogenheit und größtmögliche Repräsentativität erreicht werden.[3] Ein solches Wörterbuch will nicht allein durch die Masse der integrierten Daten beeindrucken, sondern es soll handhabbar sein, und der Benutzer sollte im Idealfall das finden, was er nachschlägt. Ein Werk, das es sich zum Ziel gesetzt hat, den Wortschatz der deutschen Gegenwartssprache repräsentativ abzubilden, ist *WAHRIG Deutsches Wörterbuch* — klassisches einbändiges Bedeutungswörterbuch mit heute rund 260.000 Stichwörtern. Die Stichwörter der Erstausgabe 1966 waren von Gerhard Wahrig noch in mühseliger Handarbeit zusammengetragen worden.

2. Lexikographie heute

Heute ist die Digitalisierung authentischer Sprachdaten ein Standard, der aus der Wörterbucharbeit nicht mehr wegzudenken ist. Die Vorteile liegen auf der Hand: Die riesigen Datenmengen aus den verschiedensten Quellen und Medien ermöglichen eine umfassende systematische Analyse, automatisierte Abgleiche sind schnell und damit ‚zeitnah' durchzuführen, und dies gewährleistet größtmögliche Aktualität. Und mit Hilfe einer einheitlichen Annotierung schließlich kann das Material relativ ‚objektiv-empirisch' ausgewertet werden. Die Korpusanalyse ist dafür das Schlüsselinstrumentarium.

Dies berührt allerdings nicht nur das bereits erwähnte Problem statistischer Frequenzanalysen. Auch hier stellt sich die Frage nach den Kriterien der Auswahl für eine qualitative Analyse und damit dem Anforderungsprofil, das an ein Korpus als Basis für

[3]Natürlich ist ‚absolute Repräsentativität' mit keinem Korpus zu erreichen; zur Relation von Größe und Repräsentativität vgl. Schierholz (2005, 7-10) und Lemnitzer/Zinsmeister (2006, 50-53).

eine Beschreibung der deutschen Gegenwartssprache gestellt wird. Die spezifische Zusammensetzung des Korpus spielt dabei eine wichtige Rolle.

2.1 Korpusbasierte Lexikographie – Anforderungsprofil

Ein solches Korpus sollte neben einem angemessenen Umfang vier weitere Kriterien erfüllen.

1. Es sollte ‚gattungsorientiert' sein, also speziell entwickelt für klassische allgemeinsprachliche Wörterbücher, und nicht primär Spezialwortschätze berücksichtigen, wie etwa ein Korpus belletristischer Literatur. Die Textsorte ‚Zeitungen und Zeitschriften', die zum größten Teil Gebrauchswortschatz enthält, bietet sich bei allen Einschränkungen dafür am ehesten an.
2. Es sollte ‚zielgruppenorientiert' aufbereitet sein, also den Sprachgebrauch eines breiten Nutzerpublikum widerspiegeln und auf wissenschaftliche und nicht wissenschaftliche Benutzer gleichermaßen ausgerichtet sein, aber z. B. auch möglichst alle Altersgruppen und verschiedene regionale Sprachspezifika abdecken.
3. Vor allem sollte es ‚anwendungsorientiert' sein für den Lexikographen, d. h. mit einer einheitlichen Annotation versehen, einer linguistischen und metalinguistischen Kodierung nach Wortart, Ressort, Stilebene, regionalem Vorkommen u. a. Erst dies schafft die Voraussetzung für Möglichkeiten etwa der Kookurrenz- und Kontextanalyse.
4. Und schließlich sollte ein solches Basiskorpus ‚gegenwartsorientiert' sein, d. h. die aktuelle deutsche Sprache betreffen und keine bzw. nur in geringem Umfang historische Texte enthalten.

3. Das WAHRIG Textkorpus^{digital}

Das WAHRIG Textkorpus^{digital} erfüllt diese Kriterien; mit mittlerweile 1,6 Milliarden Wortbelegen gehört es zu den zurzeit größten Korpora des deutschen Sprachraums.[4] Es ist vollständig und einheitlich annotiert sowie aktuell und authentisch in der Konzentration auf die zentrale Textsorte Zeitungen und Zeitschriften. Innerhalb dieser Textsorte bildet es weitgehend repräsentativen Wortschatz ab: Wichtige Zeitungen und Zeitschriften aus den verschiedensten Regionen des deutschsprachigen Raums, z. B. Österreichs und der Schweiz, sind vertreten, ebenso zentrale Fachgebiete und Sachbereiche wie (Natur-)wissenschaft oder Freizeit sowie unterschiedliche Ziel- und Altersgruppen, so etwa

[4]Die Auswertungen für die Wörterbucharbeit wurden zum allergrößten Teil aus Ergebnissen gewonnen, die mit dem WAHRIG Textkorpus^{digital} in digitalen Abgleichen ermittelt wurden. Dieses wurde in Kooperation mit der Universität des Saarlandes, Saarbrücken, speziell für die lexikographische Arbeit entwickelt. Daneben wurden bei punktuellen Analysen auch andere Korpora bzw. Quellen herangezogen, soweit öffentlich zugänglich, so z. B. das IDS-Korpus COSMAS II, das DWDS-Kernkorpus, das Korpus Deutscher Wortschatz, Leipzig sowie die Suchmaschine Google.

Jugendliche mit stark gruppenspezifischem Wortschatz.[5] Eine automatisierte, strukturierte Aufbereitung der Texte mit der Überführung in ein einheitliches Datenformat, mit differenzierter linguistischer und metalinguistischer Kodierung sowie mit Lemmatisierung der Wortbelege ist mittlerweile als Standard bei einem qualitätvollen Korpus anzusehen.

Wie kann der Lexikograph nun ein solches Korpus nutzen? Das entscheidende Kriterium, das der Korpusanalyse zugrunde liegt, hat bereits John Sinclair bei der Erarbeitung des COBUILD-Wörterbuchs mit „the strength of the corpus evidence" (Sinclair 1987, 65) bezeichnet: das der Frequenz. Wenn ein Korpus entsprechend zusammengesetzt ist, kann man daraus erkennen, welche Wörter der deutschen Sprache besonders frequent sind und welche daher in jedem Fall in einem allgemeinsprachlichen Wörterbuch verzeichnet sein sollten. Wenn diese Wörter gegen die Wortbelege der einzelnen Wörterbuchsubstanzen abgeglichen werden, ergibt sich der Grundstock, den ein Wörterbuch überhaupt ausmacht. Dabei bietet die Wörterbucharbeit den entscheidenden Vorteil, in den bestehenden Substanzen eine solide Wortauswahl als Grundlage zu haben, gegen die das Korpus abgeglichen werden kann

Eine Auflistung nach reiner Frequenz ist allerdings nur bedingt ergiebig.[6] Die lexikographische Erfahrung zeigt, dass man die Zahl der abzugleichenden Wörter sehr hoch ansetzen muss, um nach Frequenzkriterien einen repräsentativen Wortschatz ermitteln zu können. Sind diese Grenzen jedoch dann entsprechend festgelegt, ist die Tatsache, dass ein Wort im Korpus relativ häufig belegt, in der Wörterbuchsubstanz jedoch nicht zu finden ist, ein Indiz dafür, dass das Wort als Lemma aufgenommen werden sollte (‚Wortlücke'). Umgekehrt können Wörter, die im Korpus nur in sehr geringer Frequenz auftreten, unter gewissen Bedingungen ausgefiltert werden (‚Wortleiche'). Daraus ergibt sich der methodische Zugriff für die Nutzung des Korpus, bei dem verschiedene Zugänge ineinandergreifen.

---

[5]Das WAHRIG Textkorpus[digital] umfasst Wortbelege u. a. aus folgenden Medien: Berliner Zeitung, Süddeutsche Zeitung, Der SPIEGEL, Neue Zürcher Zeitung, Der Standard, Spektrum der Wissenschaft, FÜR SIE, BRAVO.

[6]Dies zeigt Carmen Scherer anhand einer Frequenzliste der IDS-Korpora. Die häufigsten Wörter sind zunächst Artikel, Konjunktionen und Präpositionen (Scherer 2006, 49).

3.1 Die Nutzung des Korpus – methodischer Zugriff

Neben der ‚absoluten Frequenz', die ohnehin nur bei der Prüfung sehr großer Datenmengen gewinnbringend ist, ist die ‚relative Frequenz' ein wichtiges Kriterium. So können etwa mittels Wortfrequenzen im Jahrgangsvergleich Neologismen aufgespürt werden, in deren Umfeld wiederum weitere Neologismen. Des Weiteren kann mit Hilfe von ‚Selektion', z. B. von Anglizismen, eine Prüfung von Schreibweisen und grammatischen Regularitäten vorgenommen werden, so etwa für die Beobachtung flektierter Formen wie *timen* oder *downloaden*. Durch eine ‚Kookkurrenzanalyse' kann ferner Sprachbeobachtung durch Ermittlung von Kollokatoren und semantischen Kontexten erfolgen. Und schließlich kann in einem ‚komparativen Verfahren' Schreibbeobachtung im Zuge der Rechtschreibreform vorgenommen werden. Insgesamt werden diese qualitativen und quantitativen Methoden in Kombination für die Erstellung und Bearbeitung von Wörterbüchern genutzt – für die zentralen Aufgaben lexikographischer Arbeit.

3.2 Korpusbasierte Lexikographie – zentrale Aufgaben

Auf der Ebene der Stichwortauswahl zentral ist die Ermittlung von ‚Lücken' und ‚Leichen' zur Substanz-Aktualisierung. Bei der Bearbeitung von Stichwortartikeln können durch Korpusanalysen semantische, orthografische oder grammatische Modifikationen erfasst werden. Einer der wichtigsten Bereiche der Wörterbucharbeit ist ferner die Suche nach Neologismen.

Ein weiterer Aspekt der Korpusarbeit, der im Zuge der Rechtschreibreform zunehmend relevant geworden ist, ist die Beobachtung des Schreibusus. Und schließlich kann man anhand des Korpus mit Hilfe der Anwendungsbeispiele eine Bestandsaufnahme von authentischem Sprachgebrauch vornehmen. Auf der Basis korpuslinguistischer Verfahren kann so eine Dokumentation zu Sprachstatus und Sprachwandel der deutschen Gegenwartssprache entstehen.

3.3 Substanzaktualisierung: fehlende Wörter

Es liegt auf der Hand, wie Wortlücken im Wörterbuch ermittelt werden können. Dies sind Wörter, die im Korpus vermehrt auftauchen, in der Wörterbuchsubstanz jedoch nicht. Andere Fragen ergeben sich durch Anfragen an die WAHRIG-Sprachberatung, etwa ob in allen morphologisch denkbaren Fällen die femininen Formen eines Stichworts verzeichnet sein sollten.[7]

---

[7]Das Korpus gibt auf diese Frage in den meisten Fällen eindeutige Antworten: *Ministerin* z. B. ist mit inzwischen 23.352 Belegen immerhin ein Viertel so häufig erfasst wie die maskuline Form, aber auch die Belegzahlen für *Managerin* (2.952 vs. 106.438) oder *Trinkerin* (140 vs. 2.107) sind ein deutliches Signal dafür, die femininen Formen als Lemmata zu verzeichnen. Anders verhält es sich mit Feminina wie z. B. *Angsthäsin* (1 vs. 234) oder *Erbsenzählerin* (3 vs. 261), die im Sprach-

3.4 Substanzaktualisierung: veraltende Wörter

Bei den ‚Wortleichen', den veralteten Wörtern, zeigt eine Korpusanalyse zu der flektierten Form *vermag* interessante Ergebnisse. Die Vorkommenshäufigkeit dieses Wortes ist von 1995-2008 auf die Hälfte gesunken. Das Wort weist zwar insgesamt immer noch eine hohe Frequenz auf, es werden jedoch deutlich Prozesse des Veraltens sichtbar (die umgekehrte Entwicklung sieht man bei *kann* in dieser Verwendung).

Abb. 1

3.5 Semantische, grammatische und orthografische Modifikationen

Auch Modifikationsprozesse auf den verschiedenen Ebenen können auf Korpusbasis im Rahmen der Substanzaktualisierung verfolgt werden. Ein Fall von ‚semantischer Modifikation' ist das Wort *schwächeln*: ein Beispiel dafür, wie ein Wort in eine andere Bedeutungsebene gelangt. Der Begriff taucht 1995 zum 1. Mal im Korpus insgesamt 17 Mal auf, und zwar im Bereich Sport mit 14 Belegen gegenüber 3 Belegen in anderen Bereichen. Ab 1999 ist ein steiler Anstieg der Frequenz bis auf 388 Belege im Jahr 2008 zu erkennen, allerdings nur noch mit 57 Belegen im Sport, 331 Belegen in anderen Bereichen, vornehmlich in der Wirtschaft. Grammatische und orthografische Modifikationen werden bei Integrations- und Grammatikalisierungsprozessen etwa bei der Konjugation von Anglizismen sichtbar, so z. B. bei der Schreibung des Partizips Perfekt von *timen* oder *downloaden* (*getimed* oder *getimt; downgeloaded* oder *gedownloadet*).

Abb. 2

3.6 Die Erweiterung der Stichwortsubstanz durch Neologismen

Neologismen sind ein Schwerpunkt der lexikographischen Arbeit, diversifizierte Methoden der Neologismenrecherche und -beobachtung damit wichtige Funktionen eines guten Textkorpus. An dem prominenten Neologismus *Wahl-O-Mat*, einer Kunstwort-Ad-hoc-Bildung, sieht man deutlich, wie Begriffshäufigkeit und Frequenz eines Wortes im Korpus gesellschaftliche Entwicklungen widerspiegeln.

Abb. 3: Neologismen zu *Wahl O-Mat*. Die Grafik zeit einen deutschen Anstieg der Wortfrequenz in Wahljahren 2002, 2005 sowie im Vorfeld der Bundestagswsahl 2009.

Das Gleiche gilt für Neologismen wie *Eliteuniversität* – in diesem Zusammenhang auch für *Exzellenzcluster* –, ebenso wie für *Abwrackprämie*, bei der es in den Jahren 2008 und

---

gebrauch nicht nachzuweisen sind.

noch stärker in der 1. Hälfte von 2009 zu einem steilen Anstieg kommt. Zu prüfen ist in diesen Fällen jeweils, ob diese Entwicklung anhält.[8]

Signifikanzen für einen typischen Neologismus zeigen die Korpusbelege des Wortes *twittern*® (= ‚über Internetportale kurze Blogs austauschen'). Dass es sich um einen Neologismus handelt, wird bereits dadurch signalisiert, dass das Wort in Anführungszeichen steht. Bei der Prüfung der Anwendungsbeispiele erhält man ferner Aussagen über den Anwendungsbereich, das Internet, aber auch Hinweise auf eine etwaige übertragene Bedeutung (Politiker twittern, doch ggf. auch Pflanzen). Und schließlich gibt es Hinweise auf andere Neologismen im Umfeld, z. B. *Blogger* oder *Follower*.

Abb. 4

All diese Beispiele spiegeln historisch-gesellschaftliche Entwicklungen. Sie zeigen, dass korpuslinguistische Verfahren es ermöglichen, Aussagen über Sprachstatus und Sprachwandel zu treffen, Integrationstendenzen bei Fremdwörtern zu signalisieren, semantische Differenzierungen und Verschiebungen anzuzeigen sowie Strukturen von Stil- und Kommunikationswandel zu beschreiben. Dies sind die traditionellen Bereiche der lexikographischen Arbeit, die bisher durch die Korpustechnologie und Korpuslinguistik abgedeckt wurden.

4. Korpusanalyse im Rat für deutsche Rechtschreibung

Ein neuer Aspekt ist erst vor kurzem in den Fokus der Korpusarbeit getreten: die Sprach- und Schreibbeobachtung für den ‚Rat für deutsche Rechtschreibung', der 2004 von der deutschen Kultusministerkonferenz als Reaktion auf die verstärkte Kritik an der Rechtschreibreform von 1996 eingesetzt worden ist. Eine seiner drei Hauptaufgaben ist die Beobachtung des Schreibgebrauchs. Diese Aufgabe wird zurzeit primär auf empirischer Ebene durch die ‚AG Korpus' wahrgenommen, die aus Mitgliedern der drei Institutionen IDS, DUDEN und WAHRIG zusammengesetzt ist und den Schreibusus der sogenannten ‚professionellen Schreiber' analysiert. Dabei stellen sich zwei grundsätzliche Fragen: Auf welche Weise und mit welchen Mitteln kann Schreibbeobachtung betrieben werden? Und in welchen Bereichen ist Schreibbeobachtung überhaupt relevant? Die Korpuslinguistik spielt dabei eine entscheidende Rolle.

4.1 Die AG Korpus: Analyse- und Auswertungskriterien

Die AG Korpus untersucht auf der Basis der Neuregelung 2006 den Schreibgebrauch nach drei Kernkomplexen:

---

[8]Dass Neologismen auch Eintagsfliegen sein können, sieht man z. B. bei einem Wort wie *Tamagotchi*®, weniger deutlich bei *Elchtest*, dessen Vorkommenshäufigkeit aber ebenfalls stark gesunken ist. Solche Wörter werden im Korpus als Streichkandidaten in regelmäßigen Abständen beobachtet.

1. Akzeptanz der Neuregelung,
2. Präferenz bei einer oder mehreren Schreibvarianten,
3. Abweichung von der Normschreibung (aufgrund von Übergeneralisierung).

4.2 Akzeptanz der Neuregelung

Bei der Analyse dieser Frage setzt die AG Korpus auf den Zeitungsjahrgängen von 1995-2008 auf: ein Jahr vor Inkrafttreten der Rechtschreibreform, als die alte Schreibung noch verbindlich war, über die erste Modifizierung der Regelung von 1996 im Jahr 2004 bis zu der grundlegenden Neuregelung vor allem der Getrennt- und Zusammenschreibung (GZS) im Jahr 2006. (Die gepunktete Kurve bezeichnet die Altschreibung, die durchgezogene die Neuschreibung; Angaben in Prozent.) Die Analysen belegen, dass Neuschreibungen in der Regel 4 Jahre brauchen, um sich zu etablieren. Exemplarisch werden hier einige paradigmatische Fälle der Laut-Buchstaben-Zuordnung (LBZ) herausgegriffen, da in diesem Bereich gegenüber 1996 keine Veränderungen vorgenommen wurden und der Akzeptanzprozess in diesem Bereich am besten beobachtet werden kann.[9]

Ein prominentes Phänomen der Rechtschreibreform im Bereich LBZ, amtliches Regelwerk, ist die neue Regel, dass in Komposita beim Aufeinanderfolgen von drei gleichen Buchstaben immer alle drei geschrieben werden. Eine Korpusanalyse zum Lemma *Schritttempo* zeigt paradigmatisch für die große Mehrzahl in dieser Fallgruppe eine Akzeptanz der Neuschreibung von nahezu 100%. Parallelanalysen einer relevanten Anzahl von Fällen verifizieren die Hypothese, dass eindeutig nachvollziehbare Regeln, die nach formalen Kriterien interpretierbar sind, stärker und schneller akzeptiert werden als solche, die von einem semantischen Kontext abhängig sind, wie es in der GZS häufig der Fall ist.

Abb. 5

Ein Gegenbeispiel findet sich im Regelkomplex ‚Volksetymologien'. Hier zeigt sich ein weniger eindeutiges Bild. Während Korpusanalysen etwa zu *belämmert* eine weitgehende Akzeptanz der volksetymologischen Deutung (von ‚Lamm') und damit auch eine Akzeptanz der Neuschreibung mit ‚ä' zeigen, weist die Bestandsaufnahme bei *Gämse* noch im Jahr 2008 nur eine Akzeptanz der Neuschreibung von 60% gegenüber 40% nicht mehr zugelassener Altschreibungen auf. Solche Fälle sind langfristig für die weitere Beobachtung vorgesehen.

Abb. 6

---

[9]Insgesamt wurden bisher rund 250 Untersuchungen in den verschiedensten Bereichen vorgenommen, systematisiert und klassifiziert. Die meisten beziehen sich auf hoch- oder mittelfrequente Wörter des Deutschen. Asterisken bezeichnen eine zu dem angegebenen Zeitpunkt nicht zulässige Schreibung.

4.3 Präferenz bei Schreibvarianten

Im Fall von Schreibvarianten stehen vor allem Fremdwörter von jeher im Fokus der Betrachtung, sind sie doch aufgrund natürlicher Integrationsprozesse von der Herkunfts- in die deutsche Sprache in besonderer Weise für zwei oder mehr Schreibvarianten prädestiniert. Am Beispiel des Lemmas *Photovoltaik* wurde die neue Regel, dass die morphologischen Bestandteile *phon, phot* und *graph* generell auch *fon, fot, graf* geschrieben werden können, getestet. Eine Korpusanalyse zeigt, dass sich die 1996 neu zugelassene integrierte Variantenschreibung 2008 noch bei nur rund 25% bewegt. Parallelanalysen belegen, dass bei überwiegend fachsprachlich gebrauchten Fremdwörtern generell die bisher mögliche und als Variante noch gültige fremdsprachliche Schreibung stark bevorzugt wird.

Abb. 7

Dies bestätigt die Analyse von *Biographie/Biografie*, bei der im Gegensatz eine deutlich stärkere Akzeptanz der neuen f-Schreibung zu verzeichnen ist — ein Indiz für die Annahme, dass das Wort in der Allgemeinsprache zunehmend stärker verankert ist als im fachsprachlichen Bereich. Diese Hypothese wird durch Analyse der Anwendungsbeispiele im Korpus gestützt.

Abb. 8

Ein letztes Beispiel für die Präferenz von Schreibvarianten ist das Wort *Mafia*, dessen integrierte Variante *Maffia* bereits vor der Rechtschreibreform zugelassen war. Die Korpusanalyse zeigt deutlich, dass die integrierte Schreibung von der Schreibgemeinschaft nicht akzeptiert wird, sie liegt konstant über die Jahre verteilt bei nur 0,5%.

4.4 Abweichung von der Normschreibung aufgrund von Übergeneralisierungen

Für den letzen Kernkomplex des Analysebereichs sei das Beispiel *Dienstagabend* herausgegriffen. Hier gibt es einen nicht sehr relevanten, aber immerhin vorhandenen Anteil nicht normgemäßer Schreibungen (dunkle durchbrochene Kurve), die zu keiner Zeit normgemäß waren: die Form *Dienstag Abend.*

Abb. 9

Für den Schreibenden kommen hier offenbar divergierende Regeln in Konflikt (vgl. dazu andere Fügungen mit Tageszeiten wie *heute Abend*). Das Phänomen des Regelkonflikts ist mit noch stärkerer Frequenz auch bei einigen Fällen der GZS festzustellen, hier muss langfristig weiter beobachtet werden, um dann ggf. eine Regelanpassung oder -modifizierung vorzunehmen.

5. Korpusgestützte Lexikographie — eine problematische Basis?

Am letztgenannten Beispiel wird jedoch nicht nur das Phänomen der Übergeneralisierung durch Abweichung von der Normschreibung deutlich, vielmehr begegnen hier auch die eingangs genannten Defizite und Unzulänglichkeiten, die Korpusanalyse-Methoden als einziges Instrument lexikographischer Arbeit aufweisen, zumindest mit einem Korpus, das zum größten Teil aus Texten ‚professioneller Schreiber' zusammengesetzt ist. Prüft man nämlich das Beispiel anhand der Suchmaschine Google, so findet man 240.000 Belege für die Fehlschreibung *Dienstag Abend* im Vergleich zu 566.000 der korrekten Zusammenschreibung, also rund 30% Falschschreibungen. Ein Korpus aus Zeitungstexten spiegelt dies nur begrenzt, da es zum großen Teil auf Korrekturprogramme aufsetzt.

Ein weiteres Defizit korpusgestützter Lexikographie ist die bereits als Kernproblem identifizierte immer noch starke Fokussierung auf das Kriterium der Frequenz; alternative Verfahren werden entwickelt. Viele interessante Neologismen sind nicht unbedingt hochfrequent. Zudem lässt sich manch interessante Quelle, wie etwa die mitgehörte Konversation von Jugendlichen im Bus, nur schwer in einem Korpus abbilden, hier wäre die Ergänzung durch ein Korpus der gesprochenen Sprache wünschenswert.

Daneben bleiben die Defizite, die sich daraus ergeben, dass bisher kein Korpus zu 100 Prozent systematisch syntaktisch und semantisch analysierbar ist. Man müsste zumindest in der Lage sein, „morphosyntaktische, distributionelle und syntaktische Informationen (Valenz) aus Texten zu extrahieren" (Haid 2005, 98). So „müssen entweder bei der Analysetiefe oder aber bei der Korrektheit bzw. Vollständigkeit der Analyse Abstriche gemacht werden" (Geyken 2004, 77). Beispielsweise ist es in vielen Fällen schwer und aufwändig, etwa bei Substantivierungen die Wortarten zu ermitteln. Generell kann die Zusammenschreibung bei Verben nicht erkannt werden, wenn ihre Einzelbestandteile nicht in Kontaktstellung vorkommen (z. B. *auseinandersetzen*).

6. Zukunftsgerichtete Lexikographie

Trotz dieser Unzulänglichkeiten bleibt das Korpus auch mit seinen mechanisch-statistischen Methoden ein zentrales Instrument moderner lexikographischer Arbeit und kann im Zusammenspiel mit den verschiedenen Quellen redaktioneller Sprachbeobachtung gewinnbringende Ergebnisse liefern, die weit über das hinausgehen, was noch vor wenigen Jahren ohne Korpora möglich war.

Dies sei hier noch einmal anhand der Schreibbeobachtung verdeutlicht. Zunächst erfolgt ein digitaler Abgleich der Wörterbuch-Substanz mit dem Korpus — bei der Erstellung und Aktualisierung der Wörterbücher. Gleichzeitig werden Datenerhebungen mit Hilfe des Korpus für den Rat für deutsche Rechtschreibung vorgenommen sowie in diesem Rahmen ein punktueller Abgleich mit zwei Referenzkorpora: dem Duden-Korpus und den IDS-

Korpora.[10] Die Schreibbeobachtung im Rat für Rechtschreibung ist damit ein wichtiger Ansatz zu übergreifender empirischer Zusammenarbeit auf Korpusebene im orthografischen Bereich. Geplant ist, dass die Ergebnisse, abgestützt durch nicht korpusbasierte wissenschaftliche Analysen von Einzelphänomenen, später in eine aktualisierte Form des amtlichen Regelwerks eingehen.

Abb. 10

Recherchiert wird auch in anderen, öffentlich zugänglichen wissenschaftlichen Korpora, geprüft werden die Ergebnisse z. T. gegen die Substanzen des Österreichischen Wörterbuchs, in einigen Fällen auch gegen die Ergebnisse in Google, weil dieses insgesamt größte ‚Korpus' überhaupt trotz seiner bekannten Beschränkungen und Unzuverlässigkeiten die größten Datenmengen und die meisten Textsorten umfasst.

In den Wörterbüchern dokumentiert WAHRIG den Schreibgebrauch und integriert in die aktualisierten Auflagen die Ergebnisse des Rats für deutsche Rechtschreibung. Dies kann z. B. in Empfehlungen in paradigmatischen Fällen für bestimmte Schreibvarianten geschehen, etwa wenn bestimmte Schreibungen nicht oder kaum verwendet werden, aber auch durch Erläuterungstexte zu verschiedenen Schreibungen in den Informationskästen des Wörterbuchs.

Das Korpus ist damit die Basisinstanz, um Sprach- und Schreibwandel zu erkennen und zu erforschen und diese Erkenntnisse in den Wörterbüchern abzubilden. Gleichzeitig ist es auch immer wieder Überprüfungsinstanz des gegenwärtigen Status und damit die Grundlage moderner allgemeinsprachlicher Wörterbücher.

Literaturverzeichnis

Quellen/Korpustexte

COSMAS II: ids-mannheim.de/cosmas2
DWDS-Kernkorpus (Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts): www.dwds.de
Korpus Deutscher Wortschatz, Leipzig: www.wortschatz.uni-leipzig.de
WAHRIG Textkorpus^{digital}: www.wahrig.de
WAHRIG Die deutsche Rechtschreibung. Hg. von der WAHRIG-Redaktion. Gütersloh /München 2009.
WAHRIG Deutsches Wörterbuch. Hg. von Renate Wahrig-Burfeind. Gütersloh/München 2008.

[10]Trotz eines hohen Grades von Vergleichbarkeit sind in einigen Fällen auch Abweichungen zwischen den drei Korpora festzustellen. Dies muss unter speziellen Bedingungen weiter analysiert werden.

Wissenschaftliche Literatur

Engelberg, Stefan/Lemnitzer, Lothar (2001): Lexikographie und Wörterbuchbenutzung. Tübingen.

Geyken, Alexander (2004): Korpora als Korrektiv für einsprachige Wörterbücher. In: Zeitschrift für Literaturwissenschaft und Linguistik 34 H 136,S. 72-100.

Haid, Ulrich (2005): Corpusbasierte Gewinnung von Daten zur Interaktion von Lexik und Grammatik: Kollokation — Distribution — Valenz. In: Friedrich Lenz/Stefan J. Schierholz (Hg.): Corpuslinguistik in Lexik und Grammatik. Tübingen, S. 97-122.

Klosa, Annette (2007): Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In: Werner Kallmeyer/Gisela Zifonun (Hg.): Sprachkorpora — Datenmengen und Erkenntnisfortschritt. Berlin/New York (=Institut für Deutsche Sprache Jahrbuch 2006), S. 105-122.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen.

Lüdeling, Anke (2007): Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In: Werner Kallmeyer/Gisela Zifonun (Hg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Berlin/New York (=Institut für Deutsche Sprache, Jahrbuch 2006), S. 28-48.

Mair, Christian (1991): Quantitative or qualitative corpus analysis? Infinital complement clauses in the Survey of English Usage corpus. In: Stig Johansson/Anna-Brita Stenström (Hg.): English Computer Corpora: Selected Papers and Research Guide. Berlin/New York, S. 67-80.

Scherer, Carmen (2006): Korpuslinguistik. Heidelberg (=Kurze Einführungen in die germanistische Linguistik 2).

Schierholz, Stefan J. (2005): Einige grundlegende Überlegungen zur Korpuslinguistik. In: Friedrich Lenz/Stefan J. Schierholz (Hg.): Corpuslinguistik in Lexik und Grammatik. Tübingen, S. 1-14.

Sinclair, John (1987): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London.

Sinclair, John (1991): Corpus Concordance Collocation. Oxford.

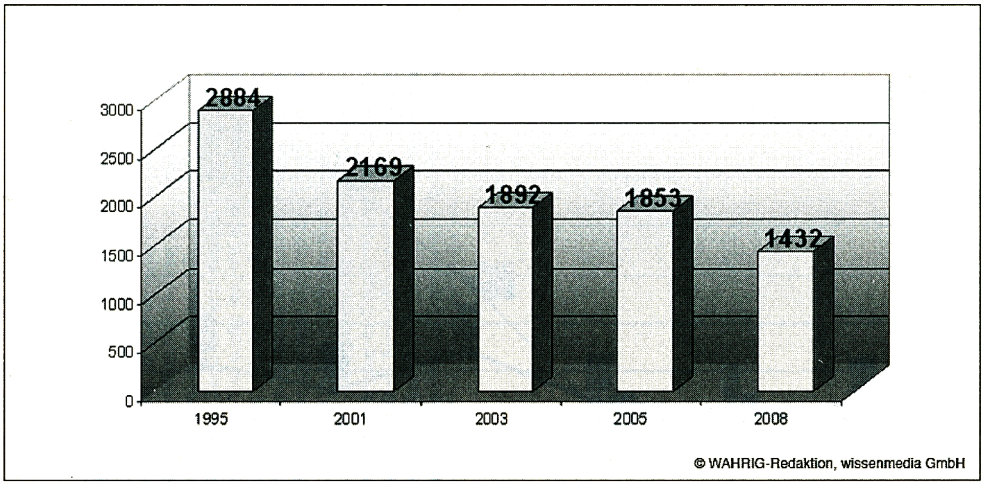


Abb. 1: Veraltende Wörter: Korpusanalyse zu *vermag*

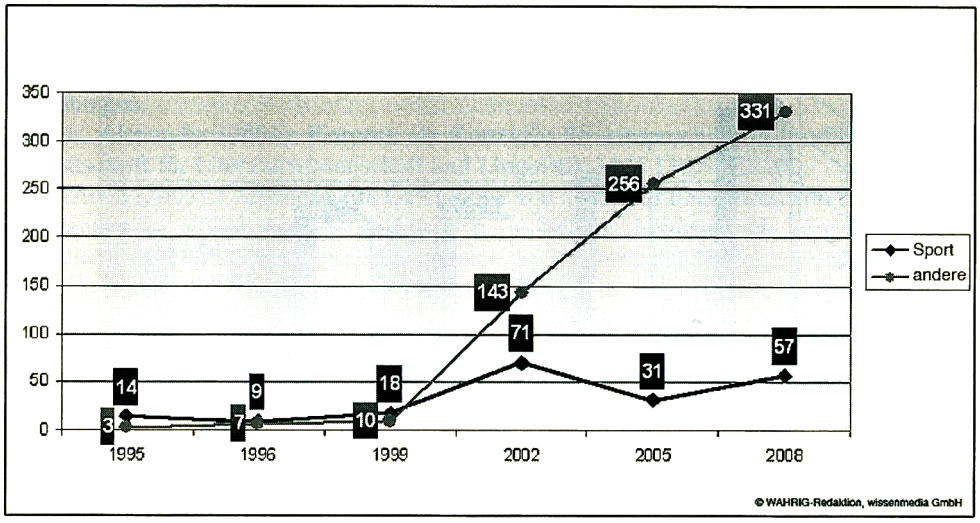


Abb. 2: Semantische Modifikationen: Korpusanalyse zu *schwächeln*

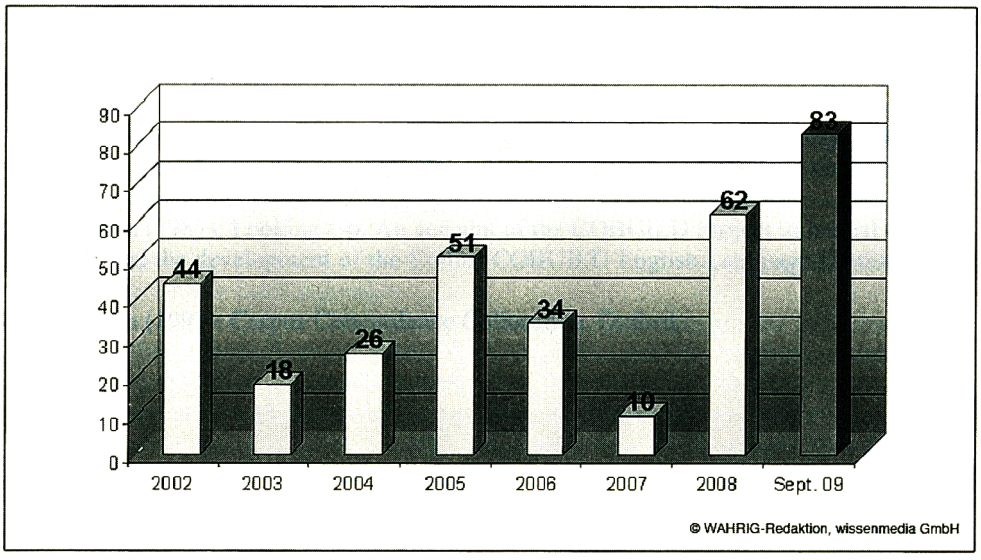
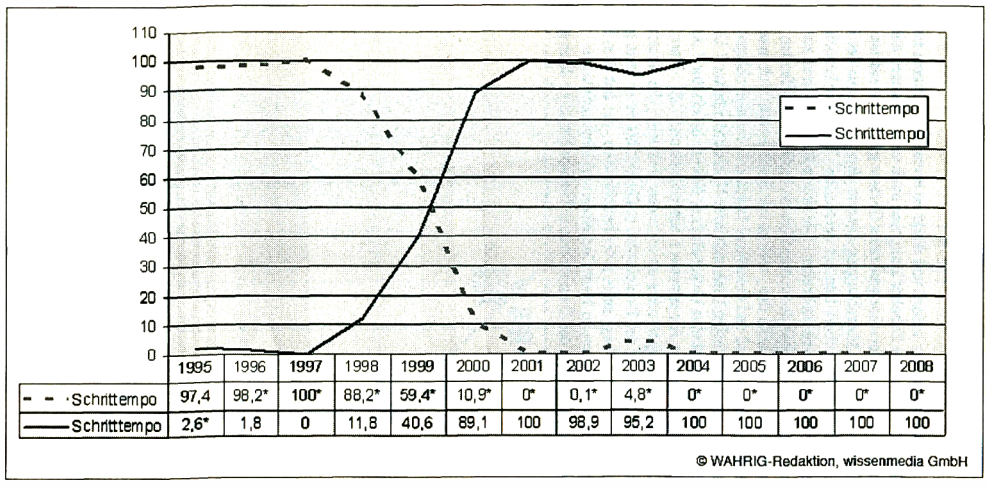


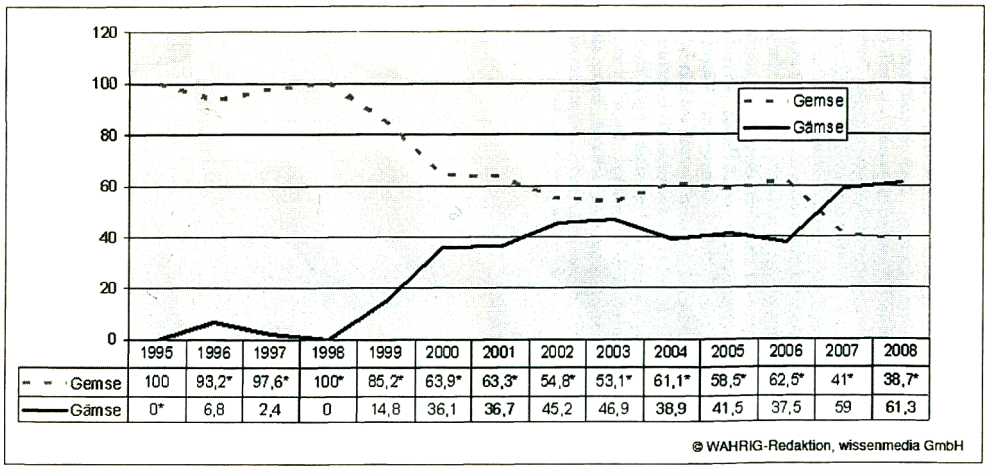
Abb. 3

Abb. 4: Neologismen zu *twittern*
Belegabfrage aus dem
WAHRIG Textkorpus[digital]



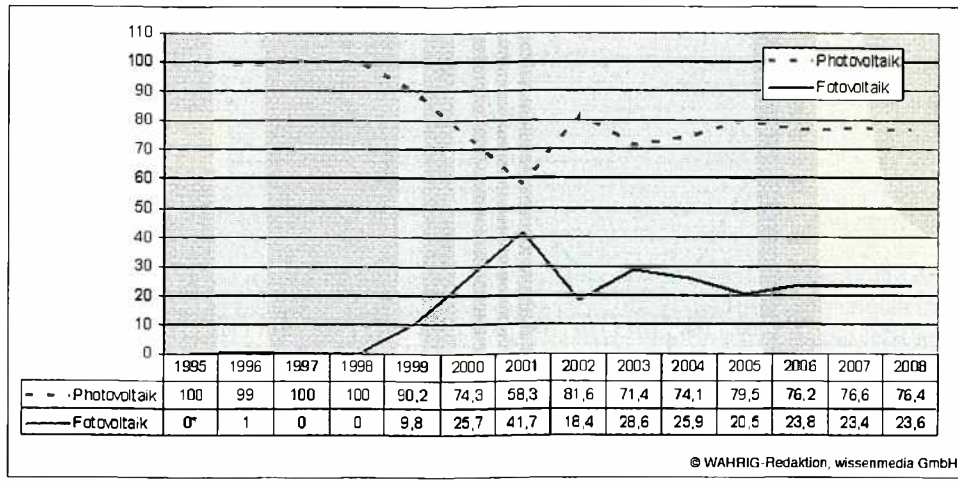
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Schrittempo	97,4	98,2*	100*	88,2*	59,4*	10,9*	0*	0,1*	4,8*	0*	0*	0*	0*	0*
Schritttempo	2,6*	1,8	0	11,8	40,6	89,1	100	98,9	95,2	100	100	100	100	100

Abb 5: Akzeptanz der Neuregelung: *Schritttempo*



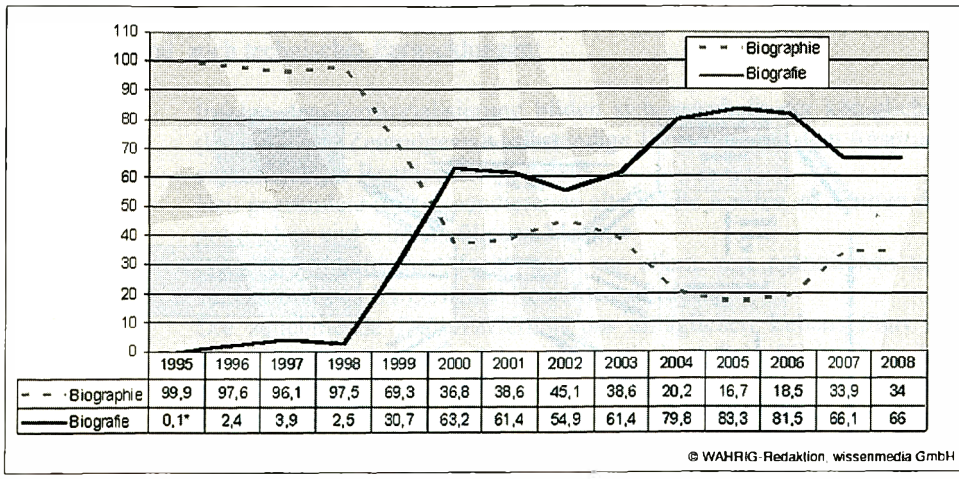
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Gemse	100	93,2*	97,6*	100*	85,2*	63,9*	63,3*	54,8*	53,1*	61,1*	58,5*	62,5*	41*	38,7*
Gämse	0*	6,8	2,4	0	14,8	36,1	36,7	45,2	46,9	38,9	41,5	37,5	59	61,3

Abb. 6: Akzeptanz der Neuregelung: *Gämse*



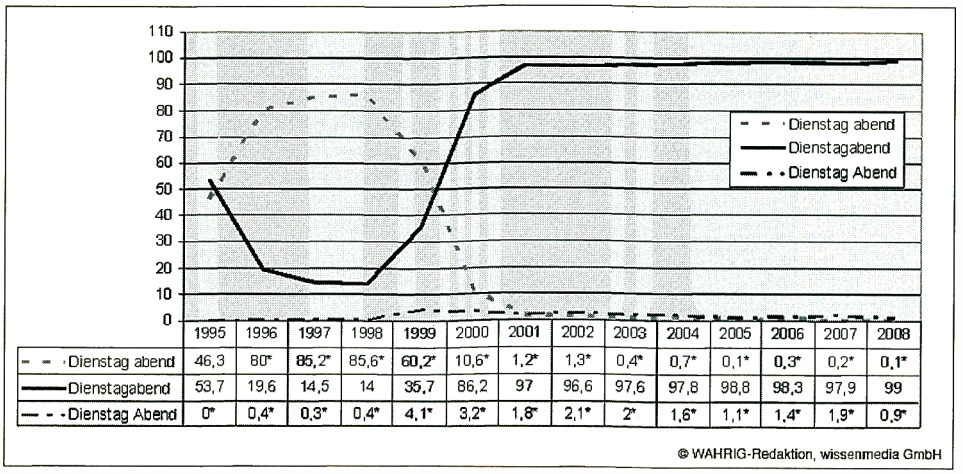
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Photovoltaik	100	99	100	100	90,2	74,3	58,3	81,6	71,4	74,1	79,5	76,2	76,6	76,4
Fotovoltaik	0*	1	0	0	9,8	25,7	41,7	18,4	28,6	25,9	20,5	23,8	23,4	23,6

Abb. 7: Präferenz bei Schreibvarianten: *Photovoltaik* vs. *Fotovoltaik*



	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Biographie	99,9	97,6	96,1	97,5	69,3	36,8	38,6	45,1	38,6	20,2	16,7	18,5	33,9	34
Biografie	0,1*	2,4	3,9	2,5	30,7	63,2	61,4	54,9	61,4	79,8	83,3	81,5	66,1	66

Abb. 8: Präferenz bei Schreibvarianten: *Biographie* vs. *Biografie*



	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Dienstag abend	46,3	80*	85,2*	85,6*	60,2*	10,6*	1,2*	1,3*	0,4*	0,7*	0,1*	0,3*	0,2*	0,1*
Dienstagabend	53,7	19,6	14,5	14	35,7	86,2	97	96,6	97,6	97,8	98,8	98,3	97,9	99
Dienstag Abend	0*	0,4*	0,3*	0,4*	4,1*	3,2*	1,8*	2,1*	2*	1,6*	1,1*	1,4*	1,9*	0,9*

Abb. 9: Abweichung von der Normschreibung aufgrund von Übergeneralisierung:*Dienstagabend*

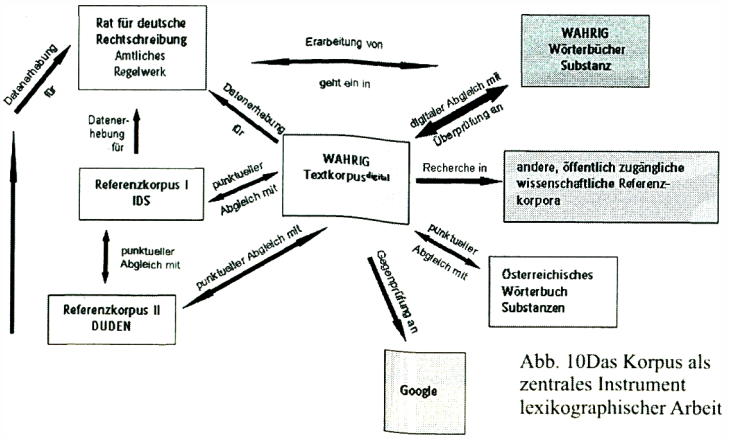


Abb. 10Das Korpus als zentrales Instrument lexikographischer Arbeit