Korpusbasierte Erstellung eines Wörterbuchs des Deutschen
Chancen und Schwierigkeiten

Magister-Hausarbeit
im Fach
„Deutsche Sprache und Literatur“

dem

Fachbereich Germanistik und Kunstwissenschaften
der Philipps-Universität Marburg

vorgelegt von
Marc Meyer
aus Kassel

Marburg 2003

Inhalt

1. Einleitung

Auf dem Wörterbuchmarkt geht der Trend dahin, dass Wörterbücher auf der Basis elektronisch vorliegender Quellen und Belege verfasst werden, dass Wörterbüchern elektronische Versionen beigegeben werden und Wörterbuchprojekte zunehmend zur Nutzung durch elektronische Medien realisiert werden.[1] Eine Ursache für diese Entwicklung liegt in der zunehmenden Verbreitung, Standardisierung, Vernetzung und Potenzierung der Leistungsfähigkeit von Computern, durch die Lexikografen vor neue Herausforderungen gestellt werden.[2] Als weitere Ursache kann gelten, dass sich in der Lexikografie ein Paradigmenwechsel vollzogen hat, seitdem computativ verfügbare Textmengen, die sich explosionsartig vermehren, durch Methoden der Korpuslinguistik den Lexikografen zugänglich werden.[3]

Das Buch ‚Corpus linguistics' von BIBER/CONRAD/REPPEN rezipiert solche Entwicklungen, wenn dort einleitend festgestellt wird, dass Online-Korpora und Analysewerkzeuge leichter zugänglich geworden sind und korpusbasierte Studien nichts Exotisches mehr an sich haben.[4] Die Autoren beklagen in diesem Zusammenhang aber auch, dass viele Forschungsberichte immer noch etwas mysteriös anmuten, da aus ihnen nicht klar hervorgeht, wie die Analysemethoden detailliert ablaufen. Dieses Problem hat sich auch beim Verfassen dieser Arbeit gestellt. Lexikografische Projekte, die auf korpuslinguistischen Methoden basieren, sind voneinander oft so verschieden, dass Studien und Forschungsberichte sehr speziell anmuten und eine große Menge Vorwissen verlangen. Weiterhin führen ganze Bücher die zentralen Stichwörter der vorliegenden Arbeit im Titel und umkreisen sie über hunderte von Seiten hinweg nur theoretisch, sodass die Erkenntnisse solcher Erörterungen oft wenig hilfreich sind, da die Konsequenzen nicht deutlich werden, die sich für die lexikografische Arbeit ergeben.

Es muss festgestellt werden, dass winzige Broschüren, Hinweise auf Internetseiten und Skripte von Vorträgen insgesamt mehr Informationsgehalt bieten als so manche große elaborierte Arbeit. Seiten im World Wide Web, die als vertrauenswürdig erachtet werden, sollen daher zitiert werden, wenngleich sich deren Inhalte kurzfristig ändern können. Die Halbwertszeit des hier erarbeiteten Wissens ist ohnehin sehr kurz anzusetzen, da Studien, die mehr als zehn Jahre zurückreichen, häufig nicht mehr als Grundlagen vermitteln können. Dies zeigt sich auch in Schwierigkeiten bei der Literaturbeschaffung selbst. Viele Publikationen

[1] Vgl. SCHMIDT/MÜLLER 2001, S. 29ff.
[2] Vgl. LENZ 2000, S. 8.
[3] Vgl. NEUMANN 1996, S.1 und SCHNEIDER 1999, S. 69.
[4] Vgl. BIBER/CONRAD/REPPEN 1998, S. IX.

finden sich weder in dem Verbundkatalog der hessischen Universitätsbibliotheken, noch in dem Verzeichnis lieferbarer Bücher und auch nicht in der Deutschen Bibliothek. Sie können nur mit viel Glück direkt bei Verlagen, Herausgebern oder als Online-Ressourcen im Internet bezogen werden.

1.1. Fragestellung

Die vorliegende Arbeit geht von einfachen Fragen aus, für die es in der wissenschaftlichen Literatur nur schwer verständliche und weit gestreute Antworten gibt. Wie werden auf der Basis elektronischer Textkorpora deutsche Wörterbücher erstellt? Welche Schwierigkeiten gibt es dabei und welche Chancen sind damit verbunden?

Bei MARTIN wird deutlich, warum zu diesen Fragen, keine schnellen Antworten zu erwarten sind, wenn dort festgestellt wird, dass es ein allgemeingültiges Verfahren zur Kompilierung von Wörterbüchern aus Korpora nicht gibt.[5] Eine weitere Schwierigkeit liegt darin, dass Fragen nach korpusbasierter Lexikografie nahezu alle Bereiche der Korpuslinguistik umfassen und Antworten fast zwangsläufig ausufern müssen, da sich hier Mittel und Objekt, Computer und Sprache in ihrer Komplexität potenzieren. Daher soll die Fragestellung fokussiert werden, indem schwerpunktmäßig nach Perspektiven und Problemen gefragt wird, die sich Lexikografie darbieten, wenn aus digitalen Quellen Wörterbücher jeder Form generiert werden sollen.

Wenn während der Betrachtungen zwangsläufig immer wieder „große“ Fragen der Linguistik in das Blickfeld treten, wird es als notwendig erachtet, die Perspektive bewusst einzuengen, da es in diesem Rahmen als nicht zweckdienlich erachtet wird, beispielsweise Metatheorien zu entwickeln, ein Sprachzeichenmodell neu zu entwerfen oder den Wortbegriff ausschweifend zu diskutieren. Fragen nach spezifischen Aspekten der Informatik oder nach der Geschichte der Korpuslinguistik müssen ebenfalls ausgeklammert werden, da sie sich zu weit von dem Gegenstand der Betrachtung entfernen.

1.2. Ziele

Diese Arbeit will den Erkenntnishorizont in Bezug auf Korpuslinguistik und Lexikografie in mehreren Punkten erweitern. Zuerst soll ein möglichst umfassender Eindruck von korpusbasierter Lexikografie in komprimierter Form vermittelt werden, damit zukünftig ein Basistext vorhanden ist, der Interessierten einen schnelleren Einstieg in die Thematik bietet

[5] Vgl. MARTIN 1995, S. 10.

als es derzeit möglich ist. Folglich richtet sich diese Arbeit nicht an Korpuslinguisten, die Lösungen zu spezifischen Einzelfragen suchen, sondern vor allem an Germanisten, die sich erstmalig für Korpuslinguistik interessieren.

Indem ein hinlänglicher Überblick darüber geboten wird, auf welche Weise das Wissen über Wörter, das in Korpora enthalten ist, extrahiert werden kann, um es dem Benutzer eines Wörterbuchs zugänglich zu machen, soll zweitens exemplarisch gezeigt werden, bei welchen Prozessen die größten Erfolge und die unangenehmsten Schwierigkeiten zu erwarten sind. Dabei sollen Basisoperationen immer wieder beispielhaft auf eine Weise dargestellt werden, die dem Betrachter weniger ihre Ergebnisse, sondern eher die Schwierigkeiten, aber auch den Einfallsreichtum jener Prozesse schildert, aus denen die Resultate gewonnen werden.

Das dritte Ziel dieser Arbeit besteht darin, den gesamten lexikografischen Prozess, den digitale Medien seit etwa 25 Jahren umwälzen, zu beschreiben, indem Korpuslinguistik nicht isoliert als Methode der Erkenntniserhebung betrachtet wird, sondern gerade auch deutlich wird, welche Chancen und Schwierigkeiten bei Kompilierung und Benutzung von Wörterbüchern entstehen, die korpusbasiert erstellt worden sind.

Schließlich sollen Linguisten in einem Ausblick darauf aufmerksam gemacht werden, welche Chancen es gibt, den Schwierigkeiten korpusbasierter Lexikografie zu begegnen, um nicht bei dem derzeit Machbaren zu verharren, sondern die Sprachwissenschaft an dieser Stelle auch anzustoßen.

1.3. Methoden

Bei MARTIN wird deutlich, dass bestimmte Basisoperationen für alle Lexikografen gleich sind, wenn es darum geht, aus dem Korpus ein Wörterbuch zu erstellen, obwohl die Techniken differieren.[6] Diese Arbeit betrachtet solche grundlegenden Verfahrensweisen korpusbasierter Lexikografie weitgehend in jener Abfolge, die tatsächlich bei der Kompilierung eines Wörterbuchs stattfindet und richtet die Perspektive auf die *„Mittel, Objekte, Benutzer und Orientierungen“*[7], die sich im Vergleich zur traditionellen Lexikografie erheblich verändert haben. Chancen und Schwierigkeiten der korpusbasierten Lexikografie werden durch den Blick auf konkrete Verfahrensweisen identifiziert, die als Lösungen für spezifische Einzelfragen vorgeschlagen werden. In diesem Sinne folgt diese Arbeit dem methodischen Pragmatismus, der gerade in der Korpuslinguistik weit verbreitet ist.

[6] Vgl. MARTIN 1995, S. 10.
[7] MARTIN 1995, S. 2.

1.4. Begriffe

Korpuslinguistik ist eine vergleichsweise junge Disziplin, deren Wurzeln im englischen Sprachraum liegen. Für viele Phänomene und Prozesse hat sich in der deutschen Literatur noch keine einheitliche Nomenklatur herausgebildet. Die englischen Fachtermini werden entweder originalgetreu wiedergegeben oder schwerfällig übersetzt und uneinheitlich flektiert. Diese Arbeit bemüht sich, Fachsprache in einer Weise zu entwickeln, die den Tendenzen gerecht werden, die sich herauszubilden scheinen.

Jenseits des Fachvokabulars, das an den entsprechenden Stellen näher spezifiziert ist, werden in dieser Arbeit immer wieder Begriffe verwendet, die Sachverhalte verkürzt darstellen. ‚Wörterbuch' und ‚Lexikon' sind hier Begriffe, die synonym in Abgrenzung zur Enzyklopädie verwendet werden. Unter ‚Korpus' wird in den meisten Fällen eine elektronisch vorliegende Volltextdatenbank in Abgrenzung zu Sammlungen von gedruckten Publikationen verstanden. Der Begriff ‚traditionelle Lexikografie' bezeichnet hier das Verfassen von gedruckten Wörterbüchern mit manuellen Methoden unter Verwendung physisch vorhandener Quellen und Belege. Hiervon abgegrenzt wird unter ‚korpusbasierter Lexikografie' das Erstellen von Wörterbüchern durch ein Verfahren verstanden, das ein elektronisches Korpus mit statistischen Methoden auswertet, um an Wissen über Wörter zu gelangen, das ohne die Rechenleistung von Computern kaum zugänglich ist.

2. Lexikologisch-lexikografische Vorüberlegungen

2.1. Lexikologische Aspekte

Lexikologie lässt sich als jenen Zweig der Linguistik definieren, der sich theoretisch mit lexikalischen Charakteristika befasst und der Lexikografie Erkenntnisse liefert, um in einem Wörterbuch Aspekte einer oder mehrerer Sprachen im Hinblick auf Benutzer zu beschreiben.[8] Korpusbasierte Lexikografie gibt nicht nur den Wörterbuchverfassern neue Werkzeuge in die Hand, sondern stellt auch neue Fragen an die Lexikologie, was beispielhaft an einer Argumentationslinie gezeigt werden soll, die NEUMANN entwickelt.[9]

Nach NEUMANN sind Wörterbücher Instrumente, die dem Linguisten morphologische Paradigmen, syntaktische Indikatoren, semantische Eigenheiten und diachrone Entwicklungen in formaler Ordnung zugänglich machen. Korpora sind für NEUMANN große Speicher nützlichen Wissens, das in den morphologischen Eigenschaften der sprachlichen Formen zum Ausdruck kommt, die wiederum im Verbund des Satzes stehen und durch die Syntax mitgeformt werden. Sie bilden ein semantisches System und repräsentieren schließlich das zu extrahierende Wissen. NEUMANN weist darauf hin, dass als entscheidende Fähigkeiten zur Gewinnung dieses Wissens Abstraktionen, Transferleistungen und Verallgemeinerungen zu nennen seien, die geeignet sind, in Korpora konkret positionierte Informationen zu einer Gesamtinformation zusammenzuführen.[10]

Wenn man wie NEUMANN davon ausgeht, dass prinzipiell alles, was es an Wissen gibt, irgendwo in einem Text realisiert ist und ein Computer alle Texte, die es gibt, verwalten kann, dann ist der Unterschied zwischen Korpora und Wörterbüchern vernachlässigbar, da er nur noch ein aspektueller Unterschied der Zugriffsart darstellt. Das Korpus kann demnach durch den geordneten Zugriff des Computers auf die einzelne Information und auf den Kontext, in dem sie steht, den Zugriff auf die verallgemeinerte Information bieten, die bisher durch den Lexikografen in einer Vielzahl von Bearbeitungen erschlossen werden muss. Bei der Extraktion des sprachlichen Wissens durch den Computer aus dem Korpus wird, im Unterschied zur traditionellen Verfahrensweise, der Kontext der zu verallgemeinernden Information berücksichtigt. Dennoch sind nach wie vor beide Formen der Zugriffsart auf sprachliches Wissen für ihre jeweils eigenen Zwecke nützlich. NEUMANNs wesentliche Erkenntnis aus seiner hier entwickelten Argumentation gipfelt in folgender Formulierung: „*Mit diesem Verständnis sind Lexika Korpora und Korpora Lexika; für Computerkorpora ist*

[8] Vgl. MARTIN 1995, S. 1.
[9] Vgl. NEUMANN 1996.
[10] Vgl. NEUMANN 1996, S. 14.

diese Differenzierung obsolet und letztlich hinderlich. Die Sprachtechnologie liefert die Methodentools, die diese Gleichung produktiv machen.“[11]

Im Hinblick auf NEUMANNS Argumentationslinie muss man sich bei der Beurteilung von Chancen und Schwierigkeiten korpusbasierter Lexikografie mehrere Dinge vor Augen führen. Erstens kann die Auswertung von Korpora mit dem Computer für den Linguisten ein wichtiges Instrument sein, um aus dem einzelnen sprachlichen Ausdruck für ein Wörterbuch verwertbare Informationen über das ganze System einer Sprache zu gewinnen. Zweitens wird man sich immer wieder vergegenwärtigen müssen, dass das elektronische Korpus gleichzeitig das Instrument der Wissenserhebung und Träger des zu erhebenden Wissens ist. Drittens muss man feststellen, dass in der Korpuslinguistik der Abstand zwischen Lexikografie und Lexikologie geringer geworden ist.[12]

2.2. Traditionelle vs. korpusbasierte Lexikografie

Zentrale Aufgabe von Lexikografie ist das Verfassen von Wörterbüchern, die sprachliche Phänomene beschreiben.[13]

Wenn man nach dem grundsätzlichen Unterschied zwischen korpusbasierter und traditioneller Lexikografie fragt, wird man wie BOGURAEV/PUSTEJOVSKY zu einem fundamentalen Problem der lexikalischen Wissenserhebung gelangen, das in dem Anspruch besteht, die Analyse des tatsächlichen Sprachgebrauchs vorzunehmen und die Lösung dieses Problems in Möglichkeiten suchen, Informationen direkt aus verwendeter Sprache heraus zu erschließen.[14] Textkorpora stellen Sprache dar, wie sie tatsächlich verwendet wird und wenn es gelingt, die Eigenschaften und Beziehungen von Wörtern mit Hilfe des Computers zu ermitteln, profitiert die Korpuslinguistik gegenüber der traditionellen Lexikografie von dem Vorteil der Unmittelbarkeit.[15] Einen weiteren Vorteil sehen BOGURAEV/PUSTEJOVSKY schlicht in dem Volumen der auswertbaren Sprache, das dem Korpuslinguisten zugänglich ist.[16] Darüber hinaus liefert der Computer sprachliche Daten komplett und verlässlich, wenn er Muster der Wortverwendung in einer größeren Spannweite und viel tiefer erfasst als es manuell möglich wäre. Schließlich ist der Sprachwandel bisher bestenfalls nach Dekaden feststellbar gewesen, wogegen dies mit Hilfe tagesaktueller Korpora theoretisch in Echtzeit erfolgen kann.

---

[11] NEUMANN 1996, S. 15.
[12] Vgl. MARTIN 1995, S. 2ff.
[13] Vgl. SCHLAEFER 2002, S. 74.
[14] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 3.
[15] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 5.
[16] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 11ff.

Die Chancen korpusbasierter Lexikografie lassen sich somit in einer ersten Skizze umreißen, auf der die direkte Erhebung sprachlichen Wissens aus tatsächlich verwendeter Sprache heraus der subjektiven Sprachkompetenz des traditionell arbeitenden Lexikografen gegenüber steht. Wenn in den folgenden Ausführungen diese Skizze überwiegend mit Schwierigkeiten korpusbasierter Lexikografie ausgefüllt wird, muss dennoch bereits festgestellt werden, dass sich das Gesamtbild eher positiv gestaltet.

3. Korpusdesign

Ein Korpus ist eine Menge von Texten, die inhaltlich oder formal als zusammengehörig betrachtet werden. SINCLAIR definiert ein Korpus als „*collection of naturally-occuring language text, chosen to characterize a state or variety of language.*"[17] SINCLAIR grenzt den Begriff ‚Korpus' an anderer Stelle von Wörtern wie ‚Archiv' oder ‚Belegsammlung' ab, da dort Auswahl und Ordnung nicht nach sprachwissenschaftlichen Kriterien erfolgen.[18] Es wird hier deutlich, wie nah sich Korpuslinguistik noch an genuinen linguistischen Leistungen der traditionellen Lexikografie bewegt, wenn es darum geht, authentisches Datenmaterial so zusammenzustellen, dass Sprache adäquat repräsentiert wird.[19] Die folgenden Betrachtungen werden zeigen, dass Korpora nicht einfach da sind, sondern zahlreiche Vorüberlegungen und Prozesse bereits in der Phase des Korpusdesigns einen wesentlichen Anteil daran haben, welche lexikalischen Erkenntnisse später extrahiert werden können.

3.1. Das Korpus zusammenstellen

3.1.1. Quellen und Belege

Traditionell arbeitende Lexikografen sind immer stark von ihrer Intuition, von ihrem Sprachgefühl bzw. von ihrer Sprachkompetenz abhängig gewesen, wenn sie mühselig zu beurteilen hatten, ob ihre Quellen den tatsächlichen Sprachgebrauch widerspiegeln und in ihrer Gesamtheit der Wörterbuchbasis genügen.[20] ENGELBERG/LEMNITZER beschreiben, welche Erleichterung die elektronische Datenverarbeitung (EDV) den Lexikografen in dem Umgang mit Quellen und Belegen verschafft, wenn sie darauf hinweisen, dass nicht nur Zettelkästen nachträglich digitalisiert worden sind, sondern ein wirklicher Fortschritt darin besteht, dass es möglich geworden ist, große Sammlungen maschinenlesbarer Texte zu erstellen und auszuwerten, um genau jene Daten zu extrahieren, die den Gebrauch eines Sprachzeichens dokumentieren.[21]

Vor dem Einsatz von Computern in der Wörterbuchproduktion sind mühevoll Quellen und Belege gesammelt, exzerpiert und für den Zettelkasten aufbereitet worden. Wie WERMKE berichtet hat die Dudenredaktion bis in die 90er Jahre hinein noch ausschließlich mit einer traditionellen Zettelkartei gearbeitet, da man sich von der selektiven Auswahl von Belegen durch Exzerptoren eine bessere Qualität versprochen hat als von einer kumulativ und damit

[17] SINCLAIR 1991, S. 171.
[18] SINCLAIR 1998, S. 114.
[19] Vgl. SINCLAIR 1998, S. 117.
[20] Vgl. ENGELBERG/LEMNITZER 2001, S. 205.
[21] Vgl. ENGELBERG/LEMNITZER 2001, S. 205.

willkürlich anwachsenden Belegmenge.[22] WERMKE identifiziert die Nachteile einer traditionellen Belegsammlung in Raumbedarf, Pflegebedarf und mangelnder Sicherheit,[23] stellt dem aber die Kosten für Verwertungsrechte, Lizenzverträge, Datenhaltung und Datensicherung[24] für ein elektronisches Volltextkorpus gegenüber. Ein wichtiger Vorteil des elektronischen Belegarchivs gegenüber dem Zettelkasten ist, dass die wenigen Felder *Stichwort, Text* und *Quellenangabe* auf der traditionellen Karteikarte der Dudenredaktion in digitalen Belegen um zahlreiche Eintragungen wie z.B. *Kontext*, *Grammatik*, *Pragmatik* ergänzt werden können.[25] Einen weiteren Vorteil des elektronischen Belegarchivs gegenüber der traditionellen Arbeitsweise sieht WERMKE auch darin, dass Textsequenzen direkt aus der Datenbank in einen Bearbeitungseditor importiert werden können, wodurch fehlerträchtiges Abschreiben entfällt.[26]

Computer gestatten folglich den Lexikografen jenseits ihrer individuellen Sprachkompetenz, die Beschreibungen zu Phänomenen der Lexik durch authentisches Sprachmaterial zu belegen und sich stärker ihrer eigentlichen Arbeit, dem Verfassen von Wörterbuchartikeln, zuzuwenden, da die zeit- und kostenintensive Beschaffung der Quelltexte erleichtert wird.[27] Die Chance, die eine solche Arbeitsweise liefert, liegt darin, dass die Verbindung zwischen den lexikografischen Beschreibungen und ihren Quellen später auf der Nutzerseite sichtbar und nachvollziehbar werden kann.[28]

3.1.2. Text

Der Begriff ‚Text' lässt sich wie bei SINCLAIR sowohl für gesprochene als auch für geschriebene Sprache verwenden.[29] Geschriebene Texte werden als lineare Abfolge von Basiseinheiten (Zeichenketten) repräsentiert. Fast jeder Text ist seit der Einführung computergesteuerter Setzmaschinen, die durch ein Satzband angewiesen werden, das sowohl den Text als auch Steuerbefehle für die Druckmaschinen enthält, in elektronischer Form vorhanden, sodass die Beschaffung von Texten rein technisch für die Korpuslinguistik keine Probleme mehr bereitet.[30] Schwierigkeiten gibt es vorwiegend mit Urheberrechten, Lizenzen

[22] Vgl. WERMKE 1998, S. 51.
[23] Vgl. WERMKE 1998, S. 51.
[24] Vgl. WERMKE 1998, S. 54.
[25] Vgl. WERMKE 1998, S. 54.
[26] Vgl. WERMKE 1998, S. 56.
[27] Vgl. STORRER 2001, S. 63.
[28] Vgl. STORRER 2001, S. 63.
[29] Vgl. SINCLAIR 1998, S. 114.
[30] Vgl. HEYN 1992, S. 2.

und mit dem Arbeitsaufwand für die Konvertierung der Daten in ein korpuskompatibles Format.

Jenseits der Beschaffungsproblematik gibt es bei der Zusammenstellung von Korpora zahlreiche Schwierigkeiten, die von der Germanistik insgesamt noch nicht hinlänglich geklärt sind, in Zusammenhang mit Text zu überwinden. Probleme im diachronen Bereich haben beispielsweise die Lexikografen des neuen MITTELHOCHDEUTSCHEN WÖRTERBUCHS, wenn die Quellen, auf die sie sich stützen, die „*ganze Breite von verhältnismäßig stark normalisierten Editionstexten bis hin zu auch graphisch überlieferungsnahen Editionen*“[31] einnehmen. In diesem Zusammenhang muss daran erinnert werden, dass geschriebene Texte sich mehr oder weniger an Schreibkonventionen halten, die einem Wandel unterliegen.

Schwierigkeiten auf der Textebene treten auch auf, wenn Strukturen in Korpora übernommen werden müssen, die nicht linear sind und sich einer eindeutigen Interpretation durch den Computer widersetzen. Beispielsweise ist es problematisch, Wörterbuchtext in ein Korpus zu importieren, wenn in den Einträgen der einzelnen Artikel die Mikrostruktur[32] teilweise typografisch dargestellt wird, da solche Typografien meist polyfunktional sind.[33] Kursive Schrift lässt sich oft erst durch den Kontext einer eindeutigen Funktion zuordnen. Hier ist demnach die Interpretationsleistung eines menschlichen Benutzers gefragt, die von einem Computer nicht erbracht werden kann.[34] Wörterbücher sind innerhalb der Artikel stark mit Verknüpfungsstrukturen durchsetzt. Es gibt Verweise, Angaben zu Belegstellen oder Verlinkungen. Solche Strukturen müssen im Korpus nachvollziehbar bleiben.[35]

Ein weiteres Problem auf der Textebene wird deutlich, wenn man sich vergegenwärtigt, dass eine große Textmenge nicht zwangsläufig den allgemeinen Sprachgebrauch dokumentiert. SINCLAIR stellt klar, dass Texte, die in Korpora importiert werden sollen, in Beziehung zur Publikumsgröße gesetzt werden müssen, da sich erst aus der Menge der Adressaten im Verhältnis zum Sprachmaterial Anhaltspunkte für die Verbreitung sprachlicher Phänomene ergibt.[36]

Mit der zunehmenden Verfügbarkeit von Text ist die Korpuslinguistik vor neue Herausforderungen gestellt, da die Zusammenstellung von Texten zukünftig nicht mehr nach praktischen Gesichtspunkten erfolgen kann, sondern nach einer genauen Typologie von Textsorten verlangt, die hinreichend verlässlich ist, spätere Analysen des Wortschatzes nicht

[31] PLATE/RECKER 2001, S. 159.
[32] Vgl. S. 57.
[33] Vgl. BÜCHEL/SCHRÖDER 2001, S. 7.
[34] Vgl. BÜCHEL/SCHRÖDER 2001, S. 8.
[35] Vgl. BÜCHEL/SCHRÖDER 2001, S. 8ff.
[36] Vgl. SINCLAIR 1998, S. 116ff.

zu verzerren. Hier wird es vor allem notwendig sein, inhaltliche Kriterien zu entwickeln, die darüber Auskunft geben, welchen Diskurs der jeweilige Text beschreibt.

3.1.3. Quantitäten

Die Korpusgröße lässt sich je nach Erkenntnisinteresse durch verschiedene Maßeinheiten darstellen. Den Informatiker interessiert vor allem der Speicherplatz, den das Korpus benötigt. Einen Archivar beschäftigt eher die Anzahl der gespeicherten Dokumente. Für die Lexikografie ist die Anzahl der enthaltenen Textwörter von besonderem Interesse.

Wenn Textwörter durch eine Suchabfrage aus dem Korpus „herausgenommen" werden, bezeichnet man sie als ‚Token'. Bei KÜNNETH werden Token definiert als „*konkrete Realisierungen einer (abstrakten) Form, dem Type. Im Falle des Wortes* linguistics *entspricht jedes Auftreten einer solchen Zeichenkette einem Token. Alle Token, die der gleichen Zeichenkette entsprechen, verweisen auf ein Type.*"[37] Hieraus ist ersichtlich, dass Token nicht zwangsläufig Textwörter darstellen, sondern Zeichenketten, die dem jeweiligen ‚Type' der Abfrage entsprechen. Bei SCHNEIDER wird deutlich, dass das Verhältnis zwischen Type und Token Auskunft zur „*lexikalischen Geschlossenheit*"[38] von Korpora gibt, die er durch eine Type-Token-Ratio (TTR) als einfache Gleichung wie folgt berechnet:

$$TTR = \frac{Anzahl\ der\ unterschiedlichen\ Wortformen}{Anzahl\ sämtlicher\ Wörter}$$

Der Wert der TTR variiert stets im Intervall von 0 bis 1, wobei die Extremwerte entweder lexikalische Monotonie (ein Type wird ständig wiederholt) oder absolute Vielfalt (jedes Token repräsentiert einen anderen Type) beziffern.

NEUMANN versucht ein exakt operationalisierbares Maß zu entwickeln, das Auskunft darüber gibt, ab welcher Größe ein Korpus hinlänglich geeignet ist, die gewünschten sprachlichen Informationen extrahieren zu können.[39] Er beschreibt ein Phänomen, das er als ‚Sättigungsgrad' bezeichnet, wenn er darstellt, dass prinzipiell jede beliebige statistisch erfassbare Eigenschaft eines Korpus berechnet und diese Berechnung nach dem Zufügen eines Textes für das umfangreichere Korpus wiederholt werden kann.[40] Nimmt die Varianz der Resultate solcher Berechnungen ab, ist das ein Hinweis für das Erreichen eines höheren Sättigungsgrad des Korpus.[41] Folglich lohnt es sich nicht mehr, ein Korpus zu erweitern,

[37] KÜNNETH 2001, S. 7 (Hervorhebungen im Original).
[38] SCHNEIDER 1999, S. 91.
[39] Vgl. NEUMANN 1996, S. 16.
[40] Vgl. NEUMANN 1996, S. 15.
[41] Vgl. NEUMANN 1996, S. 29.

wenn die Resultate für das zu berechnende Phänomen invariant werden. Bei TEUBERT findet sich ausgehend von der Erkenntnis, dass die Relation zwischen Type und Token von der Textlänge abhängig ist, eine Darstellung der Sättigung von Korpora, wenn er beschreibt, dass im ersten Satz eines Korpus noch jedes Token einem Type entspricht, in jedem weiteren Satz die Zahl der Types im Verhältnis zu den Token stetig abnimmt, bis beispielsweise die 50 Types aller Funktionswörter die Hälfte aller Token ausmachen.[42] Die Anzahl neuer Types in dem Verhältnis zu Token lässt sich demnach prinzipiell in einer Kurve beschreiben, die mit zunehmendem Text zu einer Geraden tendiert, was wiederum bedeutet, dass ein Korpus mit zunehmendem Umfang dazu neigt, sich einem konstanten Wert neuer Types anzunähern (vgl. *Abbildung 1*). Wenngleich die Quantität von Textwörtern im Korpus wesentlich ist, da in der

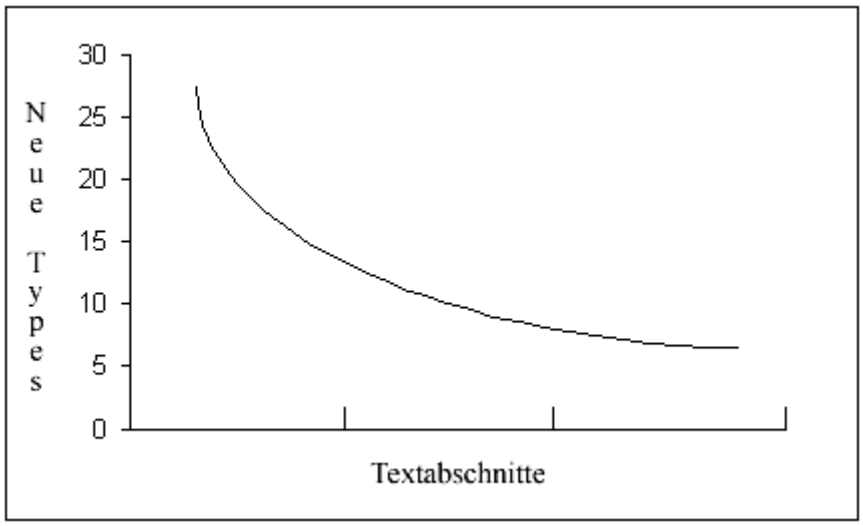


Abbildung 1. Type-Token-Relation abhängig von der Textmenge
Quelle: Frei nach TEUBERT 1998, S. 152

Auswertung großer Datenmengen der Sinn von Korpuslinguistik begründet liegt, ist es nicht zweckmäßig, bestimmte Ausmaße einzufordern, weil noch längst nicht abzusehen ist, welche Quantitäten zukünftig als Berechnungsgrundlage für welche Phänomene der Sprache herangezogen werden müssen.[43] Den meisten Forschern fällt auf, dass Korpora immer größer werden.[44] Überlegungen, ob man grundsätzlich eine möglichst große Materialsammlung anstreben sollte, führen Mitte der 90er Jahre noch zu dem Resultat, dass bereits ein Korpus von fünf bis sieben Millionen Wörtern ausreichen müsste, die Standardsprache zu dokumentieren.[45] Im Kontrast zu diesen Überlegungen stehen heute Unternehmungen wie das WORTSCHATZLEXIKON, das mehr als 500 Millionen Textwörter umfasst.[46]

[42] Vgl. TEUBERT 1998, S. 151ff.
[43] Vgl. SINCLAIR 1998, S. 116.
[44] Vgl.. MARTIN 1995, S. 7.
[45] Vgl. MARTIN 1995, S. 8.
[46] Vgl. S. 28. Vgl. auch Website des WORTSCHATZLEXIKONs http://wortschatz.uni-leipzig.de

3.1.4. Repräsentativität

Bei der Zusammenstellung von Korpora ist darauf zu achten, dass die Textauswahl so vorgenommen wird, dass die Eigenschaften von Wörtern untereinander in einem ausgewogenen Verhältnis repräsentiert werden. Bei MANNING/SCHÜTZE findet sich folgende Minimalanforderung an die Textauswahl beschrieben: „*The minimal questions we should attempt to answer when we select a corpus or report are what type of text the corpus is representative of and whether results obtained will transfer to the domain of interest.*“[47] Sie bezeichnen Korpora an gleicher Stelle als „*balanced corpora*“, wenn die darin enthaltenen Texte in Proportionen zusammengestellt sind, die im Hinblick auf das Forschungsziel vordefinierte Kriterien gewichten und in angemessenem Verhältnis berücksichtigen.[48] Aus dieser Erkenntnis folgt, dass Korpora ganz bewusst so zusammengestellt werden können, dass sie durch die Textauswahl nur einen Ausschnitt einer Sprache oder einen speziellen Diskurs dokumentieren, für den die Textsammlungen wiederum repräsentativ sein müssen.[49] Man spricht in diesem Zusammenhang von domänenspezifischen Korpora.[50]

Um den Abstand domänenspezifischer Korpora zur Standardsprache messen zu können, sind Referenzkorpora nötig, die durch repräsentative Textauswahl den sprachlichen Standard dokumentieren. Dies geschieht beispielsweise im Quellenkorpus der Dudenredaktion, indem das gesamte Textsortenspektrum des Deutschen konsultiert wird, wenn sowohl literarische Texte als auch große Tages- und Wochenzeitungen, überregionale Magazine, populärwissenschaftliche Texte und Gebrauchstexte aller Art gesammelt werden.[51] WERMKE berichtet in seinen Überlegungen zum Aufbau elektronischer Textkorpora für die Dudenredaktion, dass Jahrgangsarchive großer Zeitungen erst einmal ebenso wenig Relevanz für ein Korpus der Standardsprache haben wie elektronisch vorliegende Werkausgaben von Thomas Mann, da die Sichtung solch umfangreicher Quellen von den Exzerptoren kaum zu leisten aber dringend notwendig sei, um nicht beispielsweise in den Ergebnissen für die Suchabfrage *Fischer* überwiegend auf den deutschen Außenminister zu stoßen.[52]

TEUBERT fasst zusammen, ab wann in der Korpuslinguistik von einem standardsprachlichen Korpus gesprochen werden kann: „*Ein Referenzkorpus liegt dann vor, wenn ein Korpus möglichst viele (im Prinzip alle) als relevant erachtete Werte von möglichst vielen als*

[47] MANNING/SCHÜTZE 2000, S. 120.
[48] Vgl. MANNING/SCHÜTZE 2000, S. 120.
[49] Vgl. SINCLAIR 1998, S. 115.
[50] Vgl. VON DER GRÜN 1999, S. 5ff.
[51] Vgl. WERMKE 1998, S. 50.
[52] Vgl. WERMKE 1998, S. 53.

relevant erachteten Parametern ‚erfüllt' oder realisiert."[53] Mit dem Begriff ‚Erfüllung' bezieht sich TEUBERT an gleicher Stelle explizit auf das Prinzip der Sättigung[54] von Korpora.
Begriffe wie Repräsentativität, Balance und Ausgewogenheit sind vielfach in der korpuslinguistischen Forschung diskutiert worden. Wer in der Literatur nach exakten Parametern beispielsweise für die Zusammensetzung eines Referenzkorpus des Deutschen sucht, wird aber enttäuscht werden.

3.1.5. Datenbanken

Wenn entschieden ist, welche Texte in welchen Zusammenstellungen den Anforderungen eines lexikografischen Projekts entsprechen, müssen die Textmengen in einem Datenbanksystem gespeichert werden. Unter Datenbanksystem wird ein Dienstprogramm verstanden, das dem Nutzer die Möglichkeit gibt, auf Basis einer Datenmenge Operationen wie Einfügen, Löschen oder Suchen durchzuführen.[55] Korpusdatenbanken werden als klassische Datenbanksysteme betrachtet, da sie strukturierten Daten entsprechen, auf die sich diese Operationen anwenden lassen.[56] Da in Korpora komplexe Suchabfragen in enormen Textmengen bewältigt werden müssen, ist es nötig, die Datenstrukturen für die linguistische Datenverarbeitung zu optimieren. ZIERL beschreibt ein solches Optimierungsverfahren, das im Wesentlichen darauf beruht, Token unter Types zu subsumieren und derart zu indexieren, dass nicht mehr die konkret positionierten Zeichenketten gespeichert werden müssen, sondern lediglich Zahlenwerte, die im Binärcode so dargestellt werden können, dass sich die genaue Position jedes Token ableiten lässt.[57] Aus einer solchen Verfahrensweise resultieren weniger Bedarf an Speicherplatz und schnellere Zugriffszeiten bei der Abfrage der Datenbanken.
Das Problem, das sich hier für korpusbasierte Lexikografie stellt, liegt vor allem darin, dass die Entwicklung von Datenbanktechnologie weniger von linguistischen als von kommerziellen Interessen geleitet ist.

3.2. Arbeiten am Korpus

3.2.1. Segmentierung und Tokenisierung

Bevor Texte in die Korpusdatenbank importiert werden können, müssen ihre Zeichenketten segmentiert werden. Hierbei wird entschieden, welche sprachlichen Einheiten das Korpus

[53] TEUBERT 1998, S. 151.
[54] Vgl. S. 15.
[55] Vgl. KÜNNETH 2001, S. 16.
[56] Vgl. KÜNNETH 2001, S. 16.
[57] Vgl. ZIERL 1998, S. 39

dem Linguisten später als Token ausgeben kann. Wenn Wörter, Zahlen, Satzzeichen, Firmennamen, Abkürzungen und Internetadressen, die in einem Text enthalten sind, in Segmente unterteilt und vorhandene Trennungen rückgängig gemacht werden, spricht man daher von ‚Tokenisierung'. Praktisch gesehen geht es dabei um die schwierige Frage, was in einem nicht aufbereiteten Text als Wort gelten darf.[58] In der Linguistik wird das Wort allgemein als fundamentale sprachliche Einheit definiert, die ein Grundelement beim Aufbau von Sätzen bildet und von kompetenten Sprechern intuitiv erfassbar ist.[59] Die Korpuslinguistik geht hier wesentlich pragmatischer vor, indem sie sich nicht mit Definitionen aufhält, sondern an bereits bestehenden Konventionen orientiert oder den Wortbegriff gegebenenfalls sogar am Forschungsziel ausrichtet.[60]

ZIERL beschreibt ein Verfahren zur Tokenisierung, das hier kurz skizziert werden soll, um zu verdeutlichen, welche Schwierigkeiten bei der Segmentierung von Texten entstehen.[61] ZIERL programmiert Module, die jeweils nur einen Arbeitsschritt bewältigen, um Benutzern die Möglichkeit zu bieten, auf unterschiedliche Schreibvarianten, Formatierungen und Sprachen zu reagieren. Das erste Modul entfernt ‚whitespace characters' (Tabulatoren, Leerzeichen Zeilenumbrüche) und markiert Trennstriche, die am Ende einer Zeile stehen. Anschließend werden alle Token, die das Modul zu diesem Zeitpunkt erkennt, durch Zeilenumbrüche getrennt. Im zweiten Modul werden Zeichenketten getrennt, die durch Schrägstrich verbunden sind (z.B. *CDU/CSU*), wenn nicht eines der Wörter aus weniger als zwei Buchstaben besteht (z.B. *km/h*) oder Darstellungen von Zahlen vorliegen. Das dritte Modul trennt Satzzeichen (Komma, Semikolon, Fragezeichen, Anführungszeichen) ab und gibt sie als eigenständige Token aus, solange es sich nicht um Punkte am Ende von Wörtern handelt. Die Punktdisambiguierung wird in einem gesonderten Arbeitsschritt bewältigt. Im nächsten Modul werden die Trennungsmarkierungen analysiert, die das erste Modul markiert hat, um durch eine morphologische Analyse oder einen Abgleich mit Konkordanzlisten[62] zu verifizieren, ob es sich tatsächlich um Trennzeichen und nicht um Bindestriche oder Ergänzungszeichen handelt. Das fünfte Modul führt Zahlen zusammen, die aus mehr als drei Ziffern bestehen und mit Leerzeichen geschrieben worden sind (z.B. *20 000*). Der letzte Arbeitsgang behandelt die schwierige Disambiguierung von Punkten am Wortende, die eine Abkürzung, eine Ordinalzahl und/oder ein Satzende markieren können. Die Zeichenketten

[58] Vgl. MANNING/SCHÜTZE 2000, S. 117.
[59] Vgl. beispielsweise SCHLAEFER 2002, S. 15ff.
[60] Vgl. SCHNEIDER 1999, S. 67ff.
[61] Vgl. ZIERL 1998, S. 24ff.
[62] Vgl. S. 32.

werden zunächst mit einer Frequenzliste[63] der 200 häufigsten Abkürzungen disambiguiert. Da die Bildung von Abkürzungen produktiv ist, wird eine Liste von „Suffixen" erstellt, die nur am Ende von abgekürzten Wörtern auftreten, sodass beispielsweise durch die Endung *-tr* Abkürzungen von Straßennahmen (*Bahnhofstr., Goetherstr.)* disambiguiert werden. ZIERL stellt dar, in welchen Fällen Zeichenketten als Abkürzungen identifiziert werden:

- *„Das folgende Wort ist klein geschrieben.*
- *Das Wort besteht aus nur einem Buchstaben.*
- *Das Wort besteht aus Initialen (z.B. >>O.J.<<).*
- *Das Wort besteht nur aus Konsonantengraphemen.*
- *Das folgende Wort ist ein Satzzeichen, welches nur innerhalb eines Satzes auftreten kann.*
- *Das Wort wurde im Abkürzungslexikon gefunden.*
- *Das Ende des Wortes wurde im Suffixlexikon gefunden.*
- *Das folgende Wort ist ebenfalls eine Abkürzung.*"[64]

Wenn diese Kriterien nicht zutreffen, ist davon auszugehen, dass wahrscheinlich ein Satzende vorliegt, was sich zusätzlich mit einer Liste von 60 Artikeln, Konjunktionen und Partikeln verifizieren lässt, die nur am Anfang eines neuen Satzes groß geschrieben werden können. Weitere Anhaltspunkte für ein Satzende liefern Absatzmarkierungen und Jahreszahlen, die auf einen Punkt folgen.

ZIERL beschreibt nicht, wie sein Tokenisierer mit synthetischem Sprachbau oder bei der Interpretation von Komposita verfährt und wie er mit Wörtern umgeht, die aufgrund analytischen Sprachbaus über mehrere Graphemketten hinweg repräsentiert werden (z.B. bei der Verbflexion). Auch das Problem von Mehrwortlexemen (z.B. *kollektiver Freizeitpark*) behandelt er in seiner Arbeit nicht. Es wird entsprechend deutlich, dass allein durch Segmentierung das Problem der Wortfindung noch nicht gelöst ist. Es wird daher in einem späteren Kapitel noch einmal bei Verfahrensweisen zur Lemmatisierung[65] begegnen.

3.2.2. Annotierung

Annotierung ist prinzipiell nichts anderes als der Vorgang, der ein Korpus mit zusätzlichen Informationen anreichert.[66] Texten oder Textteilen werden Informationen über Inhalte oder Angabeklassen durch ‚Tags' angeheftet, daher wird die Annotierung auch als ‚Tagging' bezeichnet. Wenn ein Korpus mit bestimmten Kategorien annotiert ist, wird die Menge aller

[63] Vgl. S. 31.
[64] ZIERL 1998, S. 28.
[65] Vgl. S. 36.
[66] Vgl. LENZ 2001, S. 11.

verwendeten Kategorie-Etiketten ‚Tagset' genannt.[67] Im Tagset ist der Schlüssel zum Code der Annotierung festgelegt, da dort alle Tags aufgelistet und beschrieben werden, die das Korpus mit Informationen anreichern. Man unterscheidet grundsätzlich ‚Part-Of-Speech-Tagging' (POS-Tagging), wobei jedem Token beispielsweise eine Wortart zugeordnet wird, von ‚Skeletal Parsing', das zusätzlich Strukturen von Sätzen und Phrasen auszeichnet.[68]

Die morphosyntaktische Annotierung, bei der das Korpus mit Informationen zu grammatischen Kategorien wie Wortart oder Flexion angereichert wird, gilt als besonders verbreitet.[69] Vergleichsweise selten ist dagegen die syntaktische Annotierung, weil hier aufwendige manuelle Korrekturen vorgenommen werden müssen. Bei diesem ‚Parsen' werden morphosyntaktische Einheiten zu syntaktischen Einheiten zusammengefasst, sodass Textteile einem Phrasenstrukturtyp zugeordnet werden können.[70] Da auf diese Weise baumartige Strukturen beschrieben werden, die eine hierarchische Ordnung der syntaktischen Konstituenten darstellen, spricht man in diesem Zusammenhang auch von ‚Treebanking'.[71]

Kaum verbreitet ist zur Zeit noch die semantische Aufbereitung von Korpora.[72] Es wird versucht im Rahmen des Projekts ‚Semantic Information for Multifunctional Plurilingual Lexica' (SIMPLE) für 12 europäische Sprachen 20.000 Stichwörter umfassende, zum Teil korpusbasierte Lexika zu erstellen, auf deren Basis die Möglichkeit zur semantischen Annotierung der jeweiligen Sprachen eröffnet werden soll.[73]

Die Annotierung erfordert als Grundlage ein Grammatikmodell, das sich empirisch behaupten kann und daher als ‚Lexical Functional Grammar' (LFG) bezeichnet wird. Weil bereits die Wahl grammatikalischer Kategorien die späteren Forschungsergebnisse beeinflusst, muss Annotierung von Korpora immer schon als Interpretation von Texten und Sprache verstanden werden.[74]

Durch Annotierung von Korpora nehmen für Korpuslinguisten die Möglichkeiten zur Analyse sprachlicher Phänomene enorm zu, was bereits deutlich wird, wenn man sich vergegenwärtigt, dass die Disambiguierung von Homografen wie *liebe* erst gelingt, wenn man Informationen darüber hat, ob es sich bei dem Wort um ein Adjektiv oder ein Verb handelt.[75]

Für Beleg- und Quelltexte des lexikografischen Korpus wird eine besonders feinkörnige

---

[67] Vgl. KÜNNETH 2001, S. 8. Vgl auch das ‚Stuttgart-Tübingen-Tagset' auf der Website http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html
[68] Vgl. ARGENTON 1997, S. 59.
[69] Vgl. ZIERL 1998, S. 11.
[70] Vgl. LENZ 2001, S. 11.
[71] Vgl. LEZIUS 2002, S. 5.
[72] Vgl. LENZ 2001, S. 11ff.
[73] Vgl. Website von SIMPLE http://www.ub.es/gilcub/SIMPLE/simple2.html
[74] Vgl. KAMMER 1993, S. 54. Vgl. S. 21
[75] Vgl. ZIERL 1998, S. 11.

Annotierung gefordert, welche bibliografische Angaben enthält, lexikalische Einheiten auf ihre Grundform zurückführt und eindeutige syntaktische Kategorien verwendet.[76] SINCLAIR fordert, dass Korpora so annotiert werden sollten, dass es zu einem späteren Zeitpunkt wieder möglich wird, die Annotierung zu beseitigen, da noch nicht abzusehen sei, wie Korpora zukünftig annotiert werden müssten.[77]

Zusammenfassend lässt sich formulieren, dass Morphologie weitreichend, Syntax gelegentlich und Semantik fast überhaupt nicht in deutschen Korpora annotiert vorliegen.

3.2.3. Dokumenttypdefinition

Vor der Annotierung wird in einer Dokumenttypdefinition (DTD) festgelegt, welche Auszeichnungselemente es gibt, welche in anderen enthalten sein dürfen, welche Attribute sie tragen. Die Dokumenttypdefinition spezifiziert die Auszeichnungsstrukturen einer ganzen Dokumentklasse (z.B. Brief, Bedienungsanleitung, Zeitungsartikel, Wörterbuch) mittels der verwendeten Auszeichnungssprache[78] zumeist in einer gesonderten Datei[79].

Die EXPERT ADVISORY GROUP ON LANGUAGE ENGINEERING STANDARDS (EAGLES) verabschiedet auf Initiative der Europäischen Union den ‚Corpus Encoding Standard' (CES) für die Auszeichnung von Textkorpora. Der CES folgt den internationalen Richtlinien der TEXT ENCODING INITIATIVE (TEI)[80], die von EAGLES wiederum ergänzt werden müssen, da sie zu allgemein gehalten sind, um allein einen Standard zur Auszeichnung von Korpora zu bilden. Die CES-Ergänzungen beschreiben, welche Markierungen in Korpora fakultativ bzw. obligatorisch vorgenommen werden sollen. Weiterhin definieren sie präzise Werte für bestimmte Auszeichnungsattribute.[81]

Das INSTITUT FÜR DEUTSCHE SPRACHE (IDS) richtet sich bei der Auszeichnung von Korpora nach den Vorgaben von EAGLES und TEI.[82]

3.2.4. Auszeichnungssprache

Die Möglichkeiten, Texte mit Annotierungen auszuzeichnen, liegen technisch in der Kodierung durch ‚Standard Generalized Markup Language' (SGML) oder ‚eXtended Markup

[76] Vgl. STORRER 2001, S. 64.
[77] Vgl. SINCLAIR 1998, S. 120.
[78] s.u.
[79] Vgl. BURCH/FOURNIER 2001, S. 136.
[80] Vgl. S. 63.
[81] Vgl. intern 3.2.4. Auszeichnungssprache
[82] Vgl. HAß-ZUMKEHR 1998, S. 70. Vgl. auch S. 28.

Language' (XML). SGML ist bereits 1986 standardisiert (ISO 8879) worden.[83] XML (ISO-8859-1) ist genau betrachtet eine etwas strenger geregelte Ableitung von SGML, die sich vor allem im kommerziellen Bereich als Standard durchgesetzt hat.[84]

Diese Auszeichnungssprachen erfüllen nach BÜCHEL/SCHRÖDER wesentliche Anforderungen, die an die Kodierung von Texten gestellt werden müssen.[85] Zunächst gelingt es ihnen, Verweisstrukturen angemessen darzustellen. Dann sind ihre Markierungen eindeutig genug, um z.B. *zünde* von *zuende* zu unterscheiden. Sie sind durch Computerprogramme interpretierbar, die eben nicht über die Sprachkompetenz von Menschen verfügen. Ihre Nachhaltigkeit ist durch Aufwärtskompatibilität zu neuen Soft- und Hardwareentwicklungen gesichert. Ihre Strukturen erlauben die Portierbarkeit von Daten (Seitwärtskompatibilität) in die Programme verschiedenster Hersteller. BÖHME/RAHM führen weitere Argumente für die hohe Akzeptanz von XML an, wenn sie unter anderem darauf verweisen, dass die Dokumente für Menschen lesbar (kein Binärformat) und sprachenunabhängig sind, da sie den Unicode-Standard unterstützen.[86]

Wesentliches Kennzeichen der Auszeichnungssprachen SGML und XML ist, dass ihnen eine Baumstruktur zugrunde liegt. Stamm, Zweige und Blätter werden durch einen Anfangs-Tag und einen End-Tag definiert. Beispielsweise lassen sich in einem Satz Hauptwörter durch die Tags <*nomen*> und </*nomen*> in folgender Weise markieren:

<satz>
<nomen>*Tischdekorationen*</nomen>*aus*<nomen>*Plastikspinnen*</nomen>*,*
<nomen>*Gummi-Eidechsen*</nomen>*und*<nomen>*Einwegspritzen*</nomen>
bilden ein angemessen chaotisches <nomen>*Ambiente*</nomen>*.*
</satz>

In den Tags können Attribute enthalten sein, die einen bestimmten Wert beschreiben. Das folgende Beispiel zeigt die Werte *Nominativ* und *Akkusativ* für das Attribut *Kasus*, die Werte *Plural* und *Singular* für das Attribut *Numerus* sowie die Werte *Femininum* und *Neutrum* für das Attribut *Genus*:

<satz>
<nomen kasus="nom" numerus="plur" genus="fem">*Tischdekorationen*</nomen>*aus*
<nomen kasus="nom" numerus="plur" genus="fem">*Plastikspinnen*</nomen>**,**
<nomen kasus="nom" numerus="plur" genus="fem">*Gummi-Eidechsen*</nomen>*und*

[83] Vgl. BÜCHEL/SCHRÖDER 2001, S. 12.
[84] Vgl. LEZIUS 2002, S. 22. Vgl. auch Website des WORLD WIDE WEB CONSORTIUM http://www.w3.org
[85] Vgl. BÜCHEL/SCHRÖDER 2001, S. 9ff.
[86] Vgl. BÖHME/RAHM 2002, S. 1. Vgl. auch Website des Unicode-Konsortiums http://www.unicode.org

<nomen kasus="nom" numerus="plur" genus="fem">*Einwegspritzen*</nomen>
bilden ein angemessen chaotisches
<nomen kasus="akk" numerus="sing" genus="neut">*Ambiente*</nomen>.
</satz>

Verweisstrukturen sind prinzipiell ebenfalls Tags, deren Attribute auf Belegstellen führen. Das folgendes Beispiel zeigt das Wort *Homepage*, das auf eine bestimmte Website im Internet verweist:

Tourtagebücher, Fotogalerien, Rezepte und Linktips machen einen Besuch ihrer
<verweis referenz=www.rotegourmetfraktion.de>*Homepage*</verweis>
lohnenswert - empfindliche Seelen sollten der Website jedoch fern bleiben.

Die hierarchische Baumstruktur von SGML und XML wirft ein Problem auf, das nicht unerheblich ist, wenn es darum geht, Gliederungsstrukturen darzustellen, die sich überlappen.[87] Es ist beispielsweise möglich, in einem SGML-Dokument zu beschreiben, dass ein gedrucktes Wörterbuch durch einzelne Seiten gegliedert ist, die wiederum durch einzelne Artikel strukturiert sind. Wenn der Seitenumbruch aber in den Artikel gesetzt ist, kommt es zu einem Konflikt, der nur bedingt aufgehoben werden kann.[88] Eine Lösung des Problems liegt darin, mit mehreren konkurrierenden DTDs zu arbeiten. Ein solches Vorgehen hat aber zur Folge, dass der Kodierungsaufwand der Texte verdoppelt wird, da zu jeder Markierung beschrieben sein muss, auf welche DTD sie sich bezieht.
Zusammenfassend kann festgestellt werden, dass die vorhandenen Auszeichnungssprachen allen Anforderungen, die an sie gestellt werden, hinlänglich genügen und in der Wissenschaft ebenso an Akzeptanz gewinnen wie im kommerziellen Verlagswesen.

3.2.5. Automatische Annotierung

Da die manuelle Annotierung von Korpora extrem aufwendig und zeitraubend ist, versucht man Programme, die als ‚Tagger' bezeichnet werden, zu entwickeln, welche diese Aufgabe automatisch ausführen. Das Tagging läuft heute meist in zwei Phasen ab, wenn zunächst den Zeichenketten alle Tags zugeordnet werden, die als möglich gelten können und anschließend das Programm in einem zweiten Arbeitsschritt disambiguiert, welche der Möglichkeiten die wahrscheinlichere ist.[89]

[87] Vgl. BÜCHEL/SCHRÖDER 2001, S. 13.
[88] Vgl. BÜCHEL/SCHRÖDER 2001, S. 14.
[89] Vgl. ZIERL 1998, S. 11.

In der ersten Phase werden grundsätzlich Verfahrensweisen unterschieden, die an einem Trainingskorpus einen Datenabgleich vornehmen von Verfahrensweisen, die das Korpus morphologisch analysieren. Vor allem ältere Tagger[90] basieren auf der deterministischen Variante, die mit Hilfe eines Basislexikons eine Abgleichung des anzureichernden Korpus vornimmt. Die meisten Tagger beinhalten nur kleine Lexika, die größten kommen gerade auf 1.000 Wörter.[91] Solche internen Wörterbücher manuell mit syntaktischen, morphologischen und semantischen Informationen zu erstellen, ist relativ schnell und preiswert möglich. Das Basislexikon muss jedoch sorgfältig annotiert worden sein, da sonst die Gefahr besteht, dass eine hohe Fehlerquote generiert wird. Ist das Tagset des Basislexikons überdimensioniert, steigt die Fehlerquote ebenfalls, weil die Programme später Schwierigkeiten bei der Disambiguierung bekommen.[92]

Heute präferiert man Tagger, die das Korpus mit Hilfe einer LFG[93] einer morphologischen Analyse unterziehen und anschließend den Zeichenketten die entsprechenden Tags zuordnen. Hybride Systeme, welche diese Verfahrensweise mit einem Basislexikon kombinieren, werden ebenfalls angestrebt.[94]

In der zweiten Phase, wenn die Zuordnung der möglichen Tags disambiguiert wird, unterscheidet man ebenfalls zwei Verfahrensweisen, da zwischen regelbasierten und statistischen Methoden gewählt werden kann. Regelbasierte Verfahren benutzen Templates (Schablonen), welche die Annotierungen disambiguieren, indem sie die unmittelbare Umgebung der fraglichen Einheiten nach Mustern absuchen, die beispielsweise typisch für eine bestimmte Wortart sind.[95]

Statistische Verfahren arbeiten dagegen auf der Grundlage bereits annotierter Referenzkorpora, an denen sie Wahrscheinlichkeitsmodelle entwickeln, die sie wiederum mit der relativen Wahrscheinlichkeit verrechnen, dass eine Zeichenkette des Zielkorpus einer grammatischen Kategorie zuzuordnen ist.[96]

Zu beachten ist, dass auch die automatische Annotierung einen hohen Arbeitsaufwand bedeuten kann, was beispielsweise bei WILLÉE beschrieben ist, der für die Annotierung eines Korpus, das ca. eine Millionen Wörter umfasst, mit einem bereits vorhandenen Tagger einen Arbeitsaufwand von einem Mann/Jahr veranschlagt.[97] Beeindruckend ist bereits die

[90] Vgl. ZIERL 1998, S. 12.
[91] Vgl. MANNING/SCHÜTZE 2000, S. 133.
[92] Vgl. ZIERL 1998, S. 13.
[93] Vgl. ZIERL 1998, S. 12. Vgl. auch S. 27.
[94] Vgl. MARTIN 1995, S. 9.
[95] Vgl. LENDERS 1993(b), S. 374.
[96] Vgl. ZIERL 1998, S. 12ff.
[97] Vgl. WILLÉE 1993, S. 365.

Genauigkeit automatischer Annotierung, die z.B. bei LEZIUS mit 98% angegeben wird, wenn mehrere Verfahren zur Wortartbestimmung kombiniert eingesetzt werden.[98]

3.3. Ausgewählte Korpora und lexikalische Ressourcen

Die folgenden Projekte, Ressourcen und Korpora werden als einschlägige Beispiele für korpusbasierte Forschung im deutschsprachigen Raum angeführt. Dieser Überblick ist nicht vollständig. Im Entstehungsprozess dieser Arbeit ist immer wieder festgestellt worden, dass zahlreiche kleinere Lexikon-Unternehmungen offenbar korpusbasiert arbeiten, sich aber nicht entsprechend darstellen können oder einige kommerzielle Wörterbuchverlage mit Sicherheit Korpora auswerten, ihre Erkenntnisse aber aus Wettbewerbsinteresse nicht vermitteln wollen.

3.3.1. LIMAS

Das älteste deutschsprachige Korpus ist LIMAS, dessen Synchron-Schnitt 1970 gelegt worden ist. Es liegt an der Universität Bonn am INSTITUT FÜR KOMMUNIKATIONSFORSCHUNG UND PHONETIK (IKP) und enthält ca. 1,1 Millionen Textwörter bei ca. 117.000 verschiedenen Wortformen in 500 Quellen, die in ihren Textsorten dem Verhältnis des Katalogs der ‚Deutschen Nationalbibliographie' von 1970 entsprechen.[99] Das Korpus ist morphologisch annotiert und kann über die Benutzerschnittstelle COSMAS II[100] des IDS geladen und ausgewertet werden.

3.3.2. CELEX

Das LINGUISTIC DATA CONSORTIUM (LDC)[101] vertreibt eine CD-ROM des inzwischen eingestellten Projekts CELEX des DUTCH CENTRE FOR LEXICAL INFORMATION, die unter anderem ein deutsches Vollformenlexikon mit 365.530 Stichwörtern und ein Lemmalexikon mit Einträgen zu 51.728 Wortstämmen in maschinenlesbarer Form beinhaltet.[102] Der Benutzer kann hier in ca. 950 Datenbankfeldern phonologische, morphologische oder syntaktische Eigenheiten zu Wörtern kombinieren und abfragen wie z.B. Angaben über die Valenzrahmen der 9.400 eingetragenen Verblemmata.[103] CELEX basiert auf 5,4 Millionen Token, die aus

[98] Vgl. LEZIUS 2002, S. 4.
[99] Vgl. WILLÉE 1993, S. 354ff. Vgl. auch Website von LIMAS http://linux-s.ikp.uni-bonn.de/Limas/index.htm
[100] Vgl. S. 28.
[101] Vgl. Website des LINGUSITIC DATA CONSORTIUM http://www.ldc.upenn.edu
[102] Vgl. Website von CELEX http://www.kun.nl/celex/
[103] Vgl. WAUSCHKUHN 1999, S. 57ff.

Zeitungstext und Belletristik von 1949 - 1975 aus Korpora des IDS[104] stammen und auf 600.000 Token transkribierter Sprache des FREIBURGER KORPUS[105].

3.3.3. NEGRA und TIGER

Das NEGRA-Projekt ist die erste deutschsprachige Baumbank, die in 355.096 Token 20.602 syntaktisch annotierte Zeitungssätze umfasst und liegt an der Universität Saarbrücken.[106]
Tabelle 1 zeigt einen Beispielsatz des NEGRA-Korpus im Exportformat. Die ersten 60.000 Token sind bei NEGRA einer morphologischen Analyse unterzogen. In *Tabelle 1* werden die hieraus resultierenden Part-of-Speech Tags gezeigt, die dem ‚Stuttgart-Tübingen-Tagset' folgen.[107] Die ebenfalls ersichtlichen Kanten (engl. edge labels) stellen grammatikalische Funktionen dar.

%% word	tag	morph	edge	parent	secedge	comment
#BOS 1 1						
Mögen	VMFIN	3.Pl.Pres.Konj	HD	508		
Puristen	NN	Masc.Nom.Pl.*	NK	505		
aller	PIDAT	*.Gen.Pl	NK	500		
Musikbereiche	NN	Masc.Gen.Pl.*	NK	500		
auch	ADV	--	MO	508		
die	ART	Def.Fem.Akk.Sg	NK	501		
Nase	NN	Fem.Akk.Sg.*	NK	501		
rümpfen	VVINF	--	HD	506		
,	$,	--	--	0		
die	ART	Def.Fem.Nom.Sg	NK	507		
Zukunft	NN	Fem.Nom.Sg.*	NK	507		
der	ART	Def.Fem.Gen.Sg	NK	502		
Musik	NN	Fem.Gen.Sg.*	NK	502		
liegt	VVFIN	3.Sg.Pres.Ind	HD	509		
für	APPR	Akk	AC	503		
viele	PIDAT	*.Akk.Pl	NK	503		
junge	ADJA	Pos.*.Akk.Pl.St	NK	503		
Komponisten	NN	Masc.Akk.Pl.*	NK	503		
im	APPRART	Dat.Masc	AC	504		
Crossover-Stil	NN	Masc.Dat.Sg.*	NK	504		

Tabelle 1. Beispielsatz des NEGRA-Korpus
Datenquelle: NEGRA

Als Gemeinschaftsprojekt des INSTITUT FÜR GERMANISTIK in Potsdam, des INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG in Stuttgart und des DEPARTMENT OF COMPUTATIONAL LINGUISTICS AND PHONETICS in Saarbrücken liegen bei TIGER zur Zeit 40.000 Sätze

[104] Vgl. S. 28.
[105] Vgl. Website zum FREIBURGER KORPUS http://dsav-oeff.ids-mannheim.de/dsav/korpora/fr/fr_doku.htm
[106] Vgl. LEZIUS 2002, S. 8. Vgl. auch Website von NEGRA http://www.coli.uni-sb.de/sfb378/negra-corpus/
[107] Vgl. Website des INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG http://www.ims.uni-stuttgart.de

(700.000 Token) aus Zeitungstext syntaktisch annotiert vor.[108] TIGER gilt als Weiterentwicklung von NEGRA, da hier beispielsweise der Parser ANNOTATE[109] übernommen wird. Dieser Parser basiert auf einer Grammatik, die in dem Projekt PARGRAM[110] entwickelt wird, um unter anderem einen Formalismus zur Lexikonakquisition zu erzielen. An der Erarbeitung dieser LFG sind das PALO ALTO RESEARCH CENTER in Kalifornien, die Universität Stuttgart, die Universität Bergen, der Konzern FUJI XEROX in Japan, und das CENTRE FOR COMPUTATIONAL LINGUISTICS in Groß Britannien beteiligt, um eine funktionale Grammatik für die Sprachen Englisch, Französisch, Deutsch, Norwegisch, Japanisch, und Urdu zu entwickeln.[111]

3.3.4. WORTSCHATZLEXIKON

Das Projekt ‚Deutscher Wortschatz Leipzig' umfasst ca. 500 Millionen Textwörter in 35 Millionen Sätzen und ist im Internet ohne Registrierung zugänglich.[112] Hier lassen sich im WORTSCHATZLEXIKON Wortformen nachschlagen, um Angaben zu Frequenz, Grammatik, Pragmatik und Semantik zu erhalten. Besonders interessant sind die Häufigkeitsangaben zu rechten und linken Kollokaten[113] von Suchwörtern, aus denen Grafiken[114] automatisch generiert werden können, sodass ein visueller Eindruck von Bedeutungsbeziehungen entsteht.

3.3.5. IDS-Korpora und COSMAS II

Am IDS in Mannheim können über die Benutzerschnittstelle COSMAS II in 156 Korpora insgesamt ca. 1,5 Milliarden Textwörter analysiert werden.[115] Die Korpora liegen in sechs Archiven vor, die geschriebene, gesprochene, morphosyntaktisch annotierte, historische und neu akquirierte Korpora sowie phrasengegliederte „Wendekorpora" (aus der Zeit um 1989) gruppieren. Innerhalb der Archive lassen sich wiederum individuell Gruppen von Korpora zusammenstellen. COSMAS II gestattet nicht nur komplexe Suchabfragen über eine grafische Benutzeroberfläche, sondern erlaubt die Strukturierung von Massendaten zur Lemmatisierung[116], Kollokationsanalyse[117], KWIC-Anzeige[118] und zum Export der Ergebnisse.

---

[108] Vgl. Website des Projekts TIGER http://www.ims.uni-stuttgart.de/projekte/TIGER/
[109] Vgl. Website des Parsers http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html
[110] Vgl. Website des Projekts PARGRAM http://www.ims.uni-stuttgart.de/projekte/pargram/
[111] Vgl. die internationale Website des Projekts PARGRAM http://www2.parc.com/istl/groups/nltt/pargram/
[112] Vgl. Website von WORTSCHATZLEXIKON http://wortschatz.uni-leipzig.de
[113] Vgl. S. 33.
[114] Vgl. *Abbildung 4,* S. 35.
[115] Vgl. Website des IDS http://www.ids-mannheim.de/cosmas2/referenz/korpora.html
[116] Vgl. S. 36

Speziell für die lexikologisch-lexikografische Arbeit steht im IDS hausintern die COSMAS-Kookkurenzdatenbank (CCDB) zur Verfügung, in der zu 10.000 komplexen deutschen Lexemen die Ergebnisse von jeweils fünf Kookkurenzanalysen mit unterschiedlichen Parametern abgespeichert sind, wodurch bis zu 100.000 Verwendungsweisen pro Wortanalyse in Form von Hierarchien dokumentiert werden.[119] In diesem Zusammenhang ist auch eine Visualisierung von Kohäsionswerten und Kollokatorenstärke möglich.[120]

3.3.6. CISLEX

An der Universität München entsteht ein elektronisches Wörterbuch des Deutschen, das laufend mit Texten aus aktuellen Tageszeitungen und Fachzeitschriften abgeglichen wird.[121] Im Kern beinhaltet CISLEX 80.000 Einträge, die morphologische, syntaktische und semantische Informationen zu Grundformen bieten, aus denen sich Vollformen generieren lassen, was an dem folgenden Beispieleintrag deutlich wird:

*Bäckerin;fem;NS0;NP5;#12;BHA&FEM;*Pfisterin;*[122]

Durch Semikolon getrennte Felder kategorisieren hier Informationen zu Grundform (Feld 1), Genus (Feld 2), Morphologie (Felder 3-5), Semantik (Feld 6) und Synonymen (Feld 7). Aus den Feldern 3-5 lassen sich durch die kodierten Angaben zu Flexion und Kompositabildung alle Vollformen des Lemmas erzeugen. Zusätzlich bietet CISLEX Kodierungen zur Argumentstruktur von Verben und zu geografischen oder diachronen Besonderheiten von Lexemen. CISLEX dient Unternehmen, Instituten und Behörden zur automatischen Indexierung, zum Dokumentretrieval und zur Rechtschreibkorrektur.

3.3.7. DWDS

Das ‚Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts' (DWDS) ist ein Projekt der BERLIN-BRANDENBURGISCHEN AKADEMIE DER WISSENSCHAFTEN, das sich zur Aufgabe gemacht hat, den Wortschatz des 20. Jahrhunderts in einer Datenbank zu sammeln, um Benutzern über das Internet die Möglichkeit zu geben, Angaben zu Kollokationen, Phonologie und Morphologie der deutschen Gegenwartssprache online nachzuschlagen. Das Korpus beinhaltet vor allem Belletristik, journalistische Prosa, Fachprosa, Gebrauchstexte und

---

[117] Vgl. S. 33.
[118] Vgl. S. 32.
[119] Vgl. Website des IDS http://www.ids-mannheim.de Vgl. auch S. 35.
[120] Vgl. S. 34ff.
[121] Vgl. LANGER 1995, S. 2.
[122] Vgl. Website von CISLEX vom 25.08.2003 http://www.cis.uni-muenchen.de/projects/cislex.html

umfasst im Kern ca. 100 Millionen sowie in den Erweiterungen über eine Milliarde Textwörter.[123]

Das DWDS befindet sich noch im Aufbau und erlaubt zur Zeit nur eingeschränkte Suchfunktionen.

[123] Vgl. Website des DWDS http://www.dwds.de/pages/pages_info/dwds_info.htm

4. Lexikalisches Wissen extrahieren

Die Korpuslinguistik stellt nach BIBER/CONRAD/REPPEN zentrale Fragen in den Mittelpunkt ihrer Betrachtungen, die sich in etwa wie folgt zusammenfassen lassen:

- Was sind die Bedeutungen, die mit einem bestimmten Wort assoziiert werden?
- Welche Kollokationen bilden Wörter?
- Welchen außersprachlichen Bedingungen unterliegt die Auswahl eines Wortes?
- Wie hoch ist die Wahrscheinlichkeit, dass ein Wort zusammen mit anderen Wörtern in Erscheinung tritt?
- Wie sind die verschiedenen semantischen Variationen und grammatischen Funktionen verteilt?
- Wie werden scheinbare Synonyme in unterschiedlichen Umgebungen tatsächlich verwendet?[124]

Diese Fragen verdeutlichen, dass erstens ein genuines Anliegen der Korpuslinguistik die Extrahierung lexikalischen Wissens ist und zweitens, dass sich das überwiegende Interesse dabei auf die Wortsemantik richtet. Die nachfolgenden Betrachtungen nehmen die Zweiteilung dieser Erkenntnis auf, indem zunächst nach grundsätzlichen, aber auch spezifischen Verfahren der lexikalischen Wissenserhebung gefragt wird und im anschließenden Kapitel 5. das Problem der Wortbedeutung noch einmal gesondert behandelt wird.

Zuvor soll darauf hingewiesen werden, dass es kein vollautomatisches Verfahren zur lexikalischen Wissenserhebung aus Korpora gibt.[125] Je nach Forschungsinteresse muss der Lexikograf mehrere Prozesse kombinieren, um zu befriedigenden Ergebnissen zu gelangen.

4.1. Grundlegende Verfahren

4.1.1. Frequenzlisten

Frequenzlisten zählen Wörter. Sie informieren darüber, wie viele Token für mehrere Types im Korpus zu finden sind, indem den Lexemen entsprechende Ränge, Werte oder Häufigkeiten zugeordnet werden. Im Projekt ‚Deutscher Wortschatz Leipzig' wird dem Benutzer beispielsweise eine Liste der zehn häufigsten Wörter angeboten. *Tabelle 2* zeigt eine solche Liste, die hier um die Häufigkeitswerte ergänzt worden ist, mit der die jeweiligen Token zur Zeit im WORTSCHATZLEXIKON auftreten.

[124] Vgl. BIBER/CONRAD/REPPEN 1998, S. 23ff.
[125] Vgl. CLEAR 1987, S. 41.

Rang	Suchwort	Häufigkeitsklasse	Token
1	*der*	*der* ist das häufigste Wort	15.151.724
2	*die*	*der* ist ca. 2^0 mal häufiger als das Suchwort	14.548.413
3	*und*	*der* ist ca. 2^0 mal häufiger als das Suchwort	10.698.711
4	*in*	*der* ist ca. 2^1 mal häufiger als das Suchwort	7.404.753
5	*den*	*der* ist ca. 2^1 mal häufiger als das Suchwort	5.761.988
6	*von*	*der* ist ca. 2^2 mal häufiger als das Suchwort	4.562.084
7	*das*	*der* ist ca. 2^2 mal häufiger als das Suchwort	4.248.046
8	*mit*	*der* ist ca. 2^2 mal häufiger als das Suchwort	4.074.500
9	*zu*	*der* ist ca. 2^2 mal häufiger als das Suchwort	4.007.442
10	*sich*	*der* ist ca. 2^2 mal häufiger als das Suchwort	3.644.156

Tabelle 2. Frequenzliste der zehn häufigsten Wörter

Datenquelle: WORTSCHATZLEXIKON

Frequenzlisten dienen den meisten korpuslinguistischen Verfahren als Hilfsmittel. Schwierigkeiten bereiten Ergebnisse aus der Frequenzanalyse immer dann, wenn Homografen grammatisch oder semantisch nicht eindeutig interpretierbar sind.[126] Grundsätzlich ist es so, dass eine Frequenzliste nicht Wörter, sondern Wortformen nach Häufigkeit ordnet. Um Stammformen und ihnen anhängige Varianten in Frequenzlisten anzuordnen, muss zuvor eine Lemmatisierung[127] erfolgen. Frequenzlisten aus Korpora, die nicht annotiert sind, bieten folglich nur sehr begrenzten Nutzwert.[128]

4.1.2. Konkordanzlisten

Konkordanzlisten zeigen die Fundstellen eines Token, die häufig im KWIC-Format ausgegeben werden, wobei KWIC für ‚keyword in context' steht und bedeutet, dass ein Token in der Mitte einer Zeile erscheint, rechte und linke Nachbarn also sichtbar sind. Diese Darstellungsweise hat den Vorteil, dass sich dem Lexikografen sofort die situative Verwendung eines Wortes erschließt. *Abbildung 2* zeigt den Ausschnitt einer Konkordanzliste des DWDS für das Unwort des Jahres 1993 *Überfremdung*.[129] Durch die exakten Quellenangaben, die den Treffern direkt zugeordnet sind, kann sich dem Benutzer aus dieser Konkordanzliste augenblicklich erschließen, dass das Wort bereits im Nationalsozialismus häufiger Verwendung gefunden hat.

Ein Nachteil von Konkordanzlisten besteht darin, dass die Analyse der Ergebnisse weitgehend manuell erfolgen muss. Wenn die Trefferlisten zu umfangreich werden, ist ein Lexikograf auf

[126] Vgl. MANNING/SCHÜTZE 2000, S. 19.
[127] Vgl. S. 36.
[128] Vgl. BIBER/CONRAD/REPPEN 1998, S. 30.
[129] Vgl. KÜNNETH 2001, S. 17.

zusätzliche Werkzeuge angewiesen.[130] Komfortable Front-End-Lösungen wie COSMAS II erlauben dem Korpuslinguisten Abstandsoperatoren festzulegen, um die Trefferumgebung vor und hinter dem Suchwort zu begrenzen oder auszuweiten, wodurch eine tiefere Analyse der Konkordanzen ermöglicht wird.

Konkordanzlisten können beispielsweise wie beim ‚Collins COBUILD English Language Dictionary' verwendet werden, um einem Lemma Beispielsätze zuzuordnen, die aus tatsächlich verwendeter Sprache stammen.[131] Denkbar ist auch, Definitionstexte von Wörterbucheinträgen mit Hilfe von Konkordanzlisten zu gestalten.

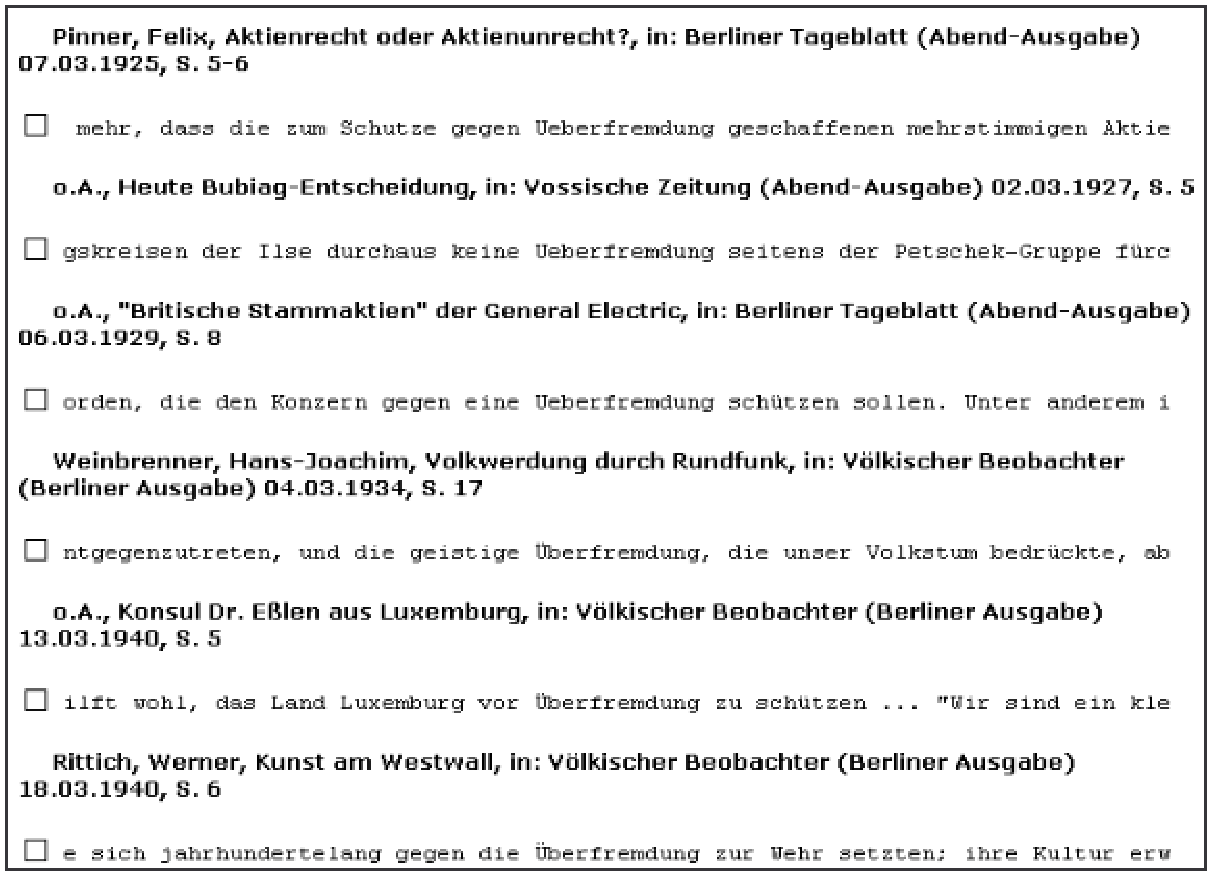


Abbildung 2. Konkordanzliste im KWIC-Format
Quelle: Frei nach DWDS

4.1.3. Kollokationen

Der Begriff ‚Kollokation' findet in der Literatur unterschiedlich Verwendung. Zunächst sind für MANNING/SCHÜTZE Kollokationen Wortverbindungen, deren „*compositionality*"[132] begrenzt ist. Darunter verstehen sie, dass normalerweise die Bedeutung einer Wortverbindung aus den Summen aller Teile addiert werden kann, während bei Kollokationen dies nur

[130] Vgl. BIBER/CONRAD/REPPEN 1998, S. 28.
[131] Vgl. ZIERL 2001, S. 16ff.
[132] MANNING/SCHÜTZE 2000, S. 151.

begrenzt möglich ist, da sie sich ein zusätzliches Bedeutungselement aneignen. Das Adjektiv *weich* addiert beispielsweise immer die gleiche sensitive Eigenschaft zu der Bedeutung von Wörtern wie *Fell* oder *Bett*. Dagegen entwickelt es in der Wortverbindung *weiche Ziele*, die 1992 als Unwort gegeißelt worden ist, zusätzlich ein euphemistisches Bedeutungselement, das sich nicht aus der Addition einzelner Wörter erschließt. Im Militärjargon wird dieser Begriff zur beschönigenden Umschreibung für das Vorhaben verwendet, gezielt die Bevölkerung und nicht die Infrastruktur eines Gegners zu vernichten.

Für die computative Lexikografie sehen MANNING/SCHÜTZE das Hauptinteresse an Kollokationen darin, die wichtigsten automatisch zu identifizieren, um sie für den Wörterbuchbenutzer in Lexikoneinträgen zugänglich zu machen.[133] Dies ist besonders sinnvoll, wenn Kollokationen häufiger in Texten auftreten als die Kollokate, aus denen sie zusammengesetzt sind.

Signifikante Kollokationen für Säuberung: ethnischen (927), ethnischer (236), Ethnische (100), Kosovo (75), Politik (62), Serben (61), Vertreibung (60), Milosevic (57), serbischen (38), zum Opfer (30), serbische (23), Krajina (21), Belgrader (19), systematischen (18), Kosten (18), begonnen (17), Reinigung (17), Strände (17), bosnischen (17), ethnische (16), Volksgruppe (16), Tschistka (16), Opfer (16), Utopie (16), Vernichtung (15), Elementen (15), Srebrenica (15), Albanern (14), Akt (14), SED (14), unterzogen (13), Vorschub (13), Volkskörpers (13), Gerd Koenen (13), Genozid (13), Rugova (13), vertrieben (13), Bosnien-Herzegowina (13), Armee (12), Kosovo-Albaner (12), Banja Luka (12), limpieza (12), Rechtsbündnis (12)
Signifikante linke Nachbarn von Säuberung: ethnischen (914), ethnischer (232), Ethnische (91), gründliche (10), Semantische (9), ethnische (8), gründlichen (7), politische (7), vollständige (7), konfessionelle (6), radikale (6), Großen (5), ideologische (5), vollständigen (5), politischer (4), regelrechte (4), soziale (4), systematischen (4), ideologischen (3), persönlich (3), regelmäßige (3)
Signifikante rechte Nachbarn von Säuberung: Grosnys (9), zum Opfer (9), unterzogen (8), Vorschub (5), ganzer (5), begonnen (4), betraf (4), betreibe (4), im Gange (4), veranlassen (4), öffentlicher (4), betrieben (3), einleiten (3), herangezogen (3)

Tabelle 3. Kollokate von Säuberung

Datenquelle: WORTSCHATZLEXIKON

Bei HEYER/QUASTHOFF/WOLFF wird ein anderer Begriff von Kollokation entwickelt, der weniger die semantische und mehr die statistische Sichtweise verfolgt. Hier wird der Begriff im Sinne von Kookkurenz verwendet, wenn sie unter Kollokationen die häufigsten linken oder rechten Nachbarn verstehen, die vor oder hinter einem Token zu finden sind.[134] Die Kookkurenzanalyse ist eine wichtige Methode zur Erschließung des sprachlichen Wissens aus Korpora. Sie geht von der Beobachtung aus, dass einige Wörter systematisch gemeinsam mit

[133] Vgl. MANNING/SCHÜTZE 2000, S. 151.
[134] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 46ff.

anderen Wörtern in Beziehung treten. Eine solche Affinität wird auch als Kohäsion bezeichnet. Kohäsionswerte können wichtige Informationen zur Identifizierung von Mehrwortlexemen, Phraseologismen, Redewendungen oder Grundbedeutungen liefern. Zwei Beispiele sollen zeigen, welche Schwierigkeiten und Chancen in Bezug auf Kollokationen für die Lexikografie zu erwarten sind.

Tabelle 3 zeigt, wie oft welche Kollokate neben dem Suchwort *Säuberung* in den Korpora des WORTSCHATZLEXIKONs zu finden sind. Die Darstellung macht deutlich, dass dieses Wort hier nicht in seiner Grundbedeutung *Reinigung* verwendet wird, sondern hauptsächlich in Verbindung mit dem Wort *ethnisch* einen Euphemismus für Völkermord bildet. Der Begriff *ethnische Säuberung*, der 1992 als Unwort des Jahres gegolten hat, prägt hier offenbar so stark die öffentliche Debatte, dass ein korpusbasiertes Verfahren nur schwer jene Grundbedeutung des Wortes erfassen kann, die der kompetente Sprecher intuitiv assoziieren würde.

Das Unwort *Ich-AG* des Jahres 2002 erschließt sich dem kompetenten Sprecher dagegen nicht intuitiv, solange er nicht durch den entsprechenden Diskurs darüber informiert ist, dass dieser Begriff einen Arbeitslosen bezeichnet, der in die Selbständigkeit gedrängt wird. Einer Visualisierung der Kollokate nach ihrer Kohäsionsstärke durch das WORTSCHATZLEXIKON (Vgl. *Abbildung 4*) gelingt es dagegen, den Bedeutungsbereich des Wortes so darzustellen, dass die Grundbedeutung des Lexems durch die Schlagworte des gesamten politischen Diskurses abgebildet wird.

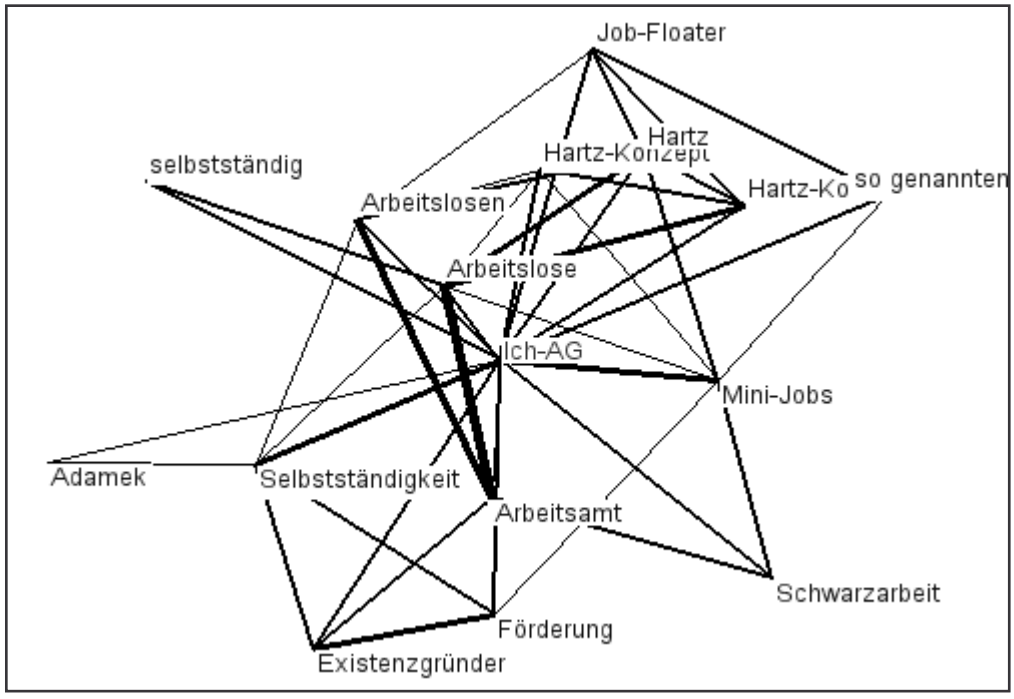


Abbildung 4. Visualisierung der Bedeutungsbeziehungen von Ich-AG

Quelle: WORTSCHATZLEXIKON

Die hier gezeigte Visualisierung entsteht im Prinzip dadurch, dass ein Netzdiagramm mit den häufigsten Kollokaten des Suchwortes erstellt wird. Anschließend werden die Kollokate befragt, in welchen Kohäsionsstärken sie untereinander vernetzt sind.
Kollokationen stehen zur Zeit im Zentrum korpuslinguistischer Interessen, da man sich von ihnen Aussagen zur Wortsemantik erhofft, die in der traditionellen Lexikografie lange Zeit nicht getroffen werden konnten.

4.1.4. Lemmatisierung

Wenn aus Korpora eine lexikalische Ressource generiert werden soll, ist es nötig, eine Lemmaliste zu erstellen, um zu entscheiden, welchen Stichwörtern die segmentierten Wortformen zugeordnet werden müssen. Diese Lemmata sind zumeist die Infinitive der Verben, die 1. Pers. Sing. der Hauptwörter, unflektierte Adjektive und Derivationen oder auch Wortbildungsmorpheme und in speziellen Wörterbüchern zuweilen die Vollformen selbst.
Mit Hilfe von Frequenzlisten, die aus großen Referenzkorpora gewonnen werden, lassen sich sprachliche Einheiten auswählen, die in den oberen Frequenzrängen den Kernwortschatz darstellen und in den unteren Rängen seltene oder fachspezifische Diskurse widerspiegeln.[135]
Zur Definition eines Kernbereiches der Sprache bieten aus Korpora extrahierte Lemmalisten auf diese Weise wichtige Entscheidungshilfen, wenngleich es nicht ratsam ist, bei der Erstellung von Stichwortlisten für Wörterbücher allein auf Frequenzlisten zu vertrauen.[136]
Eine Schwierigkeit bei der automatischen Lemmatisierung liegt in der richtigen Zuordnung von flektierten Formen und Derivationen zu den Grundformen der Stichwortlisten. Eine Möglichkeit zur Problemlösung besteht darin, durch semantische oder morphologische Disambiguierung ein System entscheiden zu lassen, ob eine finite Form wie *lag* der Grundform *liegen* oder dem Infinitiv *legen* zuzuordnen ist. Eine solche Disambiguierung kann über eine Analyse der Kollokationen erfolgen, die im WORTSCHATZLEXIKON für *legen*, *liegen* und *lag* beispielsweise in den folgenden Häufigkeiten ermittelt werden:

Kollokate von *legen*

auf den Tisch (745), *an den Tag* (670), *auf den Boden* (497), *den Grundstein für* (472), *in den Weg* (222), *Karten auf den Tisch* (211), *Steine in den Weg* (179)

Kollokate von *liegen*

auf dem Tisch (2008), *auf der Hand* (1904), *im Trend* (1181), *am Herzen* (790), *am Boden* (494)

[135] Vgl. ENGELBERG/LEMNITZER 2001, S. 208.
[136] Vgl. ENGELBERG/LEMNITZER 2001, S. 212.

Kollokate von *lag*

am Boden (1019), *im Minus* (976), *im Bett* (734), *am Herzen* (532)

Hier wird schnell deutlich, dass die Wortform *lag* ähnlich wie die Grundform *liegen* häufig im Zusammenhang mit Dativen und niemals in Verbindung mit Akkusativen auftritt, während dies bei *legen* genau umgekehrt der Fall ist. Ein automatisches System hätte demnach anhand der Kollokate einen verlässlichen Hinweis gewonnen, dass *lag* der Grundform *liegen* zuzuordnen ist, da es nach dem gleichen Kasus verlangt.

Eine weitere Möglichkeit, Wortformen ihren Lemmata zuzuführen, liegt im so genannten ‚Stemming', das relativ verbreitet ist, wobei Formen unter Stichworte subsumiert werden, indem man z.B. Morpheme oder Suffixe ignoriert.[137] Für das Lemma *abfackeln* lassen sich in COSMAS II Wortformen finden, indem eine Datenbankabfrage mit Trunkierungen (Platzhaltern) gestaltet wird. Die Suchanfrage *ab??fackel** findet in den IDS-Korpora im Archiv der geschriebenen Sprache beispielsweise die folgenden Wortformen in den jeweiligen Häufigkeiten: *abgefackelt* (215), *abzufackeln* (37), *abgefackelten* (16), *abgefackelte* (10), *Abgefackelt* (6), *abgefackeltes* (5), *abgefackelter* (3), *abgefackeltem* (1). Zunächst scheint eine solche Abfrage der Wortformen erfolgreich zu sein. Problematisch ist hier aber, dass große Schwierigkeiten entstehen, wenn versucht wird, dem Lemma *abfackeln* jene Verbformen zuzuführen, die das Präfix abgetrennt haben, um es in der Wortstellung des Satzes frei anzuordnen.

Eine dritte Möglichkeit Aussagen zur Zugehörigkeit von Wortformen und Grundformen zu treffen, liegt in der Auswertung so genannter Bigramme, die sich auf die Methode stützt, paarweise Zeichenketten miteinander zu vergleichen, bis sich aus der Anzahl der Operationen, die nötig sind, ein Wort in ein anderes zu überführen, ein Ähnlichkeitswert errechnet.[138] Dieses Distanzmaß operiert mit vordefinierten Werten für den Aufwand an Einfügungen, Auslassungen und Ersetzungen von Graphemen, die vorgenommen werden müssten, um beispielsweise ein Wort wie *abgefackelt* in *abfackeln* zu transformieren. Hier wären zwei Grapheme zu streichen und eines auszutauschen, sodass insgesamt drei Operationen benötigt werden, um zu der Grundform zu gelangen. Unter der Voraussetzung, dass die verglichenen Zeichenketten von minimal bestimmter Länge sind, lassen sich auf diese Weise, sobald der Ähnlichkeitswert eine bestimmte Schwelle überschreitet, Annahmen zur Wortverwandtschaft formulieren.

[137] Vgl. MANNING/SCHÜTZE 2000, S. 132.
[138] Vgl. SCHNEIDER 1999, S. 111.

Schwierigkeiten gibt es mit Mehrwortlexemen, da diese in den Frequenzlisten wie mehrere Spracheinheiten behandelt werden. Nach ENGELBERG/LEMNITZER funktionieren die meisten Verfahren zur Lemmatisierung von Mehrwortlexemen so, dass man einen Wert ermittelt, wie oft zwei Wörter nebeneinander aufzufinden wären, wenn man alle Wörter des Korpus zufällig verteilen würde. Anschließend berechnet man die Abweichung der tatsächlichen Wortpositionen von denen der ermittelten Zufallsverteilungen und prüft jene Wortpaare genauer, die hohe Abweichungswerte erreichen, da diese auf eine höhere Affinität der Wörter schließen lassen.[139]

MANNING/SCHÜTZE beklagen, dass der Nutzwert automatischer Lemmatisierungen nicht unbedingt größer ist als die Fehlermenge, die sie produzieren[140]. Bei LEZIUS wird die Fehlerrate dieser Verfahren mit 5% beziffert.[141] Es ist in jedem Fall davon auszugehen, dass Lemmatisierung nicht durch ein einzelnes dieser Verfahren geschehen kann, sondern Methoden kombinieren muss.[142]

4.2. Spezifische Verfahren zur lexikalischen Wissensextraktion

Seitdem es durch korpusbasierte Verfahren möglich wird, Wissen über Wörter zu sammeln, das mit traditionellen Mitteln lange nicht entdeckt werden konnte, entstehen zahlreiche Methoden, die untereinander stark variieren. Die Ansätze zur Wissenserhebung sind ebenso vielfältig und verschieden wie die Erkenntnisse, die aus ihnen gewonnen werden. Daher können hier nur einige Verfahren beschrieben werden, die exemplarisch illustrieren, auf welche Weise der Lexikografie Chancen durch den Computereinsatz geboten werden.

4.2.1. Textsorten und Domänen

Wenn sich ein Korpus aus sehr unterschiedlichen Quellen zusammensetzt, besteht die Möglichkeit, die situative Verwendung eines Wortes zu erforschen, weil davon auszugehen ist, dass ein Diskurs durch eine bestimmte Verwendungsweise spezieller Wörter in spezifischen Quellen kodiert ist.

Wichtig ist hier zunächst, dass die absolute Zahl der Treffer für die Suchabfrage eines Wortes in den verschiedenen Textsorten innerhalb eines Korpus noch keine Auskunft über die Sprachkonventionen innerhalb eines Genres gibt. Die Treffer müssen in ein Verhältnis zum

[139] Vgl. ENGELBERG/LEMNITZER 2001, S. 209.
[140] Vgl. MANNING/SCHÜTZE 2000, S. 132.
[141] LEZIUS 2002, S. 4.
[142] Vgl. SCHNEIDER 1999, S. 118ff.

Anteil der Textsorten im Korpus gesetzt werden, indem beispielsweise eine Aussage über das Erscheinen eines Wortes in einer Umgebung von 100.000 Textwörtern erfragt wird.

Bei BIBER/CONRAD/REPPEN ist ein Verfahren beschrieben, das sich verschiedener Textsorten bedient, um Synonymie zu beurteilen.[143] Sie gehen von der Beobachtung aus, dass Wörter oft intuitiv als gleichbedeutend angesehen werden, obwohl sie es nicht sind. Ihnen fällt beispielsweise auf, dass die englischen Wörter *large*, *tall* und *big* als synonym gelten, obwohl sie jeweils in ganz unterschiedlichen Kontexten verwendet werden. In der Belletristik erscheint nach ihrer Zählung das Wort *big* in einer Million Textwörter 408 mal, in Forschungsliteratur dagegen nur 31 mal. Dagegen wird das Wort *large* in belletristischen Texten nur 232 mal in einer Million Textwörtern gefunden, während es in Forschungsliteratur 605 mal auftritt.[144] Diesen Häufigkeitswerten lässt sich entnehmen, dass diese scheinbaren Synonyme Ziel gerichtet in bestimmten inhaltlichen Zusammenhängen verwendet werden. Die nähere Betrachtung der Konkordanzen von *big* ergibt dann auch, dass das Wort in Forschungsliteratur und Belletristik hauptsächlich gewählt wird, um physische Größe als absolute Eigenschaft zu beschreiben und ähnlich wie ein Farbadjektiv verwendet wird, während *large* oft in Zusammenhang mit Zahlen und Quantitäten gebraucht wird, um Proportionen zu beschreiben.[145] Die Textsorten, in denen Lexeme in bestimmten Verbreitungsmustern vorliegen, erteilen demnach wichtige Auskünfte über die Grundbedeutungen von Wörtern.

Wenn ein domänenspezifisches Korpus vorliegt oder es die Korpusstruktur gestattet, Gruppen von Texten auszuwerten, ist eine weitergehende Analyse möglich. Die Gruppierung der Quelltexte ist problematisch, da sie sowohl nach inhaltlichen als auch nach formalen Kriterien kategorisiert werden können. Es ließen sich beispielsweise gedruckte Sportberichte und Börsennachrichten inhaltlich verschiedenen Sachgebieten zuordnen oder formal gemeinsam als Zeitungstext klassifizieren. Viele Korpuslinguisten gehen auch hier sehr pragmatisch vor, da sie sich an den Vorgaben ausrichten, die ihnen die Korpusstruktur auferlegt.[146]

LJUNG untersucht, welche Unterschiede im Wortschatz auftreten, wenn über verschiedene Genre (Sport, Wirtschaft, Kunst, etc.) in Zeitungen berichtet wird und verfährt, indem er die Frequenzlisten der genrespezifischen Texte an einem Referenzkorpus abgleicht.[147] Jene Lexeme, die ungewöhnlich häufig in den oberen Frequenzrängen genrespezifischer Korpora zu finden sind, können dann als typischer Wortschatz identifiziert werden.

---

[143] Vgl. BIBER/CONRAD/REPPEN 1998, S. 65.
[144] Vgl. BIBER/CONRAD/REPPEN 1998, S. 44.
[145] Vgl. BIBER/CONRAD/REPPEN 1998, S. 45ff.
[146] Vgl. LJUNG 2002, S. 186.
[147] Vgl. LJUNG 2002, S. 183.

SCHNEIDER formuliert hierzu eine Hypothese: „*Je häufiger ein Wort in Texten einer eingeschränkten Kategorie verwendet wird und je mehr morphologische und graphemische Varianten es besitzt, umso größer ist die Wahrscheinlichkeit, daß das Wort domänenspezifisches Wissen repräsentiert.*“[148]

Es entscheidet also nicht allein die Frequenz, mit der Lexeme im domänenspezifischen Korpus im Vergleich zur Standardsprache auftreten, ob sie einem genrespezifischen Wortschatz zuzurechnen sind. Eine Aussage über die Variationsbreite der Graphemketten, die als Kandidaten in Frage kommen, kann ebenfalls Anhaltspunkte dafür bieten, ob es sich um Lexeme handelt, die eine Affinität zu bestimmten Textsorten oder Domänen zeigen.

Hierbei muss auch berücksichtigt werden, dass domänenspezifische Korpora ihre eigenen „Gesetze“ haben. SCHNEIDER versucht beispielsweise aus einer Sammlung von Geschäftsberichten den domänenspezifischen Wortschatz durch gewichtete Frequenzlisten herauszufiltern.[149] Die Gewichtung dient auch hier dazu, den wichtigsten domänenspezifischen Wörtern obere Ränge zuzuweisen, um sie später isolieren zu können.[150] Hier würde eine Rangliste aufgrund bloßer Aufsummierung der Frequenzen von Wortformen und Stammformen nicht berücksichtigen, dass Stammformen in diesem speziellen Korpus generell frequenter vorliegen als Wortformen. SCHNEIDER versucht daher, die Frequenzlisten so zu gewichten, dass in den oberen Rängen tatsächlich domänenspezifische Wörter erscheinen.

Rang	**Stammform**	**Frequenz Stammform**	**Frequenz Varianten**	**Rang nur Stammform**	**Rang nur Varianten**
1	geschäftsbericht	87	24	8	27
2	ihre	131	5	1	8
3	unternehmen	42	13	17	35
4	jahresabschlüsse	31	17	24	43
5	bitte	91	4	6	10
6	zusendung	48	7	14	31
7	meiner	37	8	19	40
8	ihnen	71	4	10	15
9	senden	31	9	24	32
10	unserer	34	7	21	36

Tabelle 4. Auszug einer gewichteten Frequenzliste

Datenquelle: SCHNEIDER 1999, S. 131

Dies gelingt ihm, indem er die Häufigkeit einer Stammform mit der Anzahl ihrer Varianten multipliziert. In *Tabelle 4* ist eine gewichtete Frequenzliste SCHNEIDERs dargestellt. Da hier

[148] SCHNEIDER 1999, S. 130.
[149] Vgl. SCHNEIDER 1999.
[150] Vgl. SCHNEIDER 1999, S. 130.

auch angegeben ist, welche Ränge und Frequenzen die einzelnen Stammformen und Varianten ohne Gewichtung einnehmen würden, kann betrachtet werden, wie das domänenspezifische Lexem *Geschäftsbericht* erst durch die Gewichtung jenen Rang in der Liste einnimmt, dem man ihm intuitiv für dieses Korpus zugestehen würde.
Zusammenfassend muss festgestellt werden, dass die Analyse von Textsorten und Domänen ein wichtiges Arbeitsgebiet korpusbasierter Lexikografie ist. Chancen bestehen hier nicht bloß in der Ermittlung domänenspezifischer Wortschätze, sondern auch in der Analyse von Grundbedeutungen des Allgemeinwortschatzes.

4.2.2. Monitoring

Manche Korpora lassen sich in Teilen so datieren, dass statistische Phänomene sprachlicher Eigenheiten in einer Zeitachse darstellbar werden, wodurch die diachrone Struktur des Sprachwandels in der Diskussion von Kurven fassbar wird.[151] TEUBERT definiert ein Monitorkorpus als „*ein Korpus, das in regelmäßigen Abständen unter möglichst genauer Beibehaltung der Kompositionsparameter ergänzt wird.*“[152] Für NEUMANN ist es beispielsweise denkbar, ein Zeitungsarchiv als Monitorkorpus zu verwenden, um darzustellen, ab wann ein Wort wie stark in Mode kommt und anschließend wieder schwindet.[153] Zu dieser Überlegung soll hier ein Experiment durchgeführt werden, das sich der Frage zuwendet, wie sich ein ‚Unwort des Jahres’ vor und nach der Ächtung durch die GESELLSCHAFT FÜR DEUTSCHE SPRACHE im Sprachgebrauch von Zeitungstexten verteilt.
Das größte online verfügbare Zeitungsarchiv im deutschen Sprachraum bilden die GENIOS-Wirtschaftsdatenbanken.[154] Hier ist es möglich, in einer Gruppe von 24 deutschsprachigen Zeitungstiteln[155] über eine Zeitspanne von drei Jahren hinweg ein Suchwort monatlich aufzufinden. Eine Suchanfrage für das Wort *Topterrorist*, das nach Ablauf des Jahres 2001 zum Unwort erklärt worden ist, ergibt für die Archivbestände vom 01.06.2000 bis 30.06.2003, die schätzungsweise 100 Millionen Textwörter[156] enthalten, eine Liste von 1.394 Treffern in 662 Zeitungsdokumenten. In den Treffern sind auch Wortformen enthalten, da mit Trunkierungen (*Topterrorist$$*) gesucht wird. Die Treffer lassen sich im Kontext betrachten,

[151] Vgl. NEUMANN 1996, S. 16.
[152] TEUBERT 1998, S. 159.
[153] Vgl. NEUMANN 1996, S. 16.
[154] Vgl. Website von GENIOS http://www.genios.de
[155] Handelsblatt, Financial Times Deutschland, Börsen-Zeitung, Süddeutsche Zeitung, Frankfurter Rundschau, Neue Zürcher Zeitung, Der Tagesspiegel, Die Tageszeitung, Die Welt, Welt am Sonntag, Die Zeit, VDI Nachrichten, Der Spiegel, Focus, Wirtschaftswoche, Börse Online, FOCUS-Money, Stern, Capital, brand eins, DMEuro, Impulse, Manager Magazin, Junge Karriere.
[156] Die Schätzung beruht auf der Abfrage einer Frequenzliste der zehn häufigsten Wörter des Deutschen.

sodass neben der quantitativen Analyse auch inhaltliche Aspekte bei der Auswertung berücksichtigt werden können.

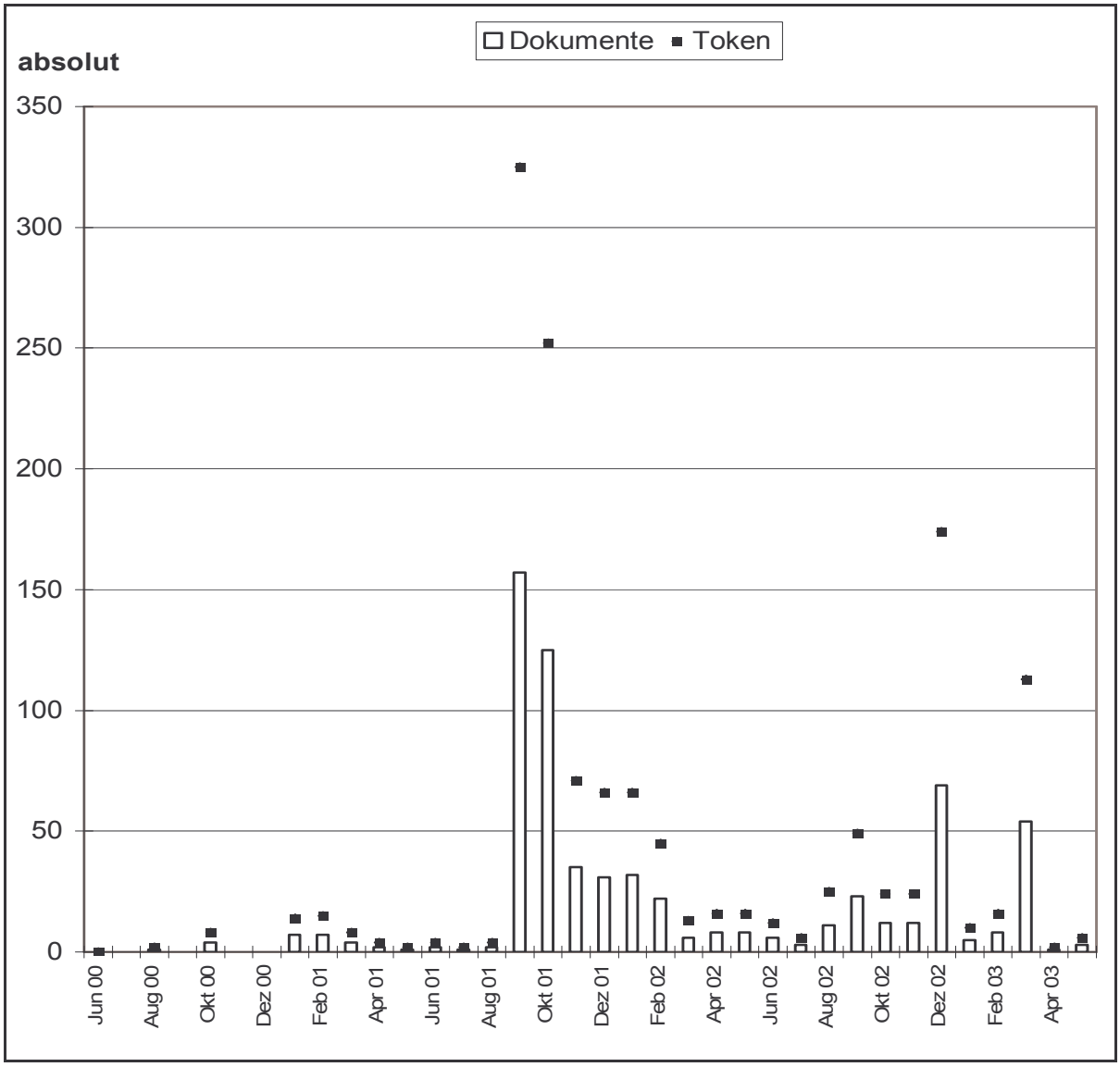


Abbildung 3. Frequenzen des Suchwortes Topterrorist *in Zeitungsdokumenten*
Datenquelle: GENIOS

Vor den terroristischen Anschlägen in New York und Washington wird *Topterrorist* 15 Monate lang lediglich in 31 Zeitungsartikeln 63 mal erwähnt. Im September 2001 kann man es in Verbindung mit den Attentaten plötzlich in 157 Dokumenten 325 mal finden. Im darauf folgenden Monat wird es noch 252 mal in 157 Zeitungsartikeln verwendet. In den anschließenden zwei Monaten sinkt die Verwendung des Wortes auf 71 Nennungen im November 2001 und 66 Nennungen im Dezember 2001. Als das Wort im Januar 2002 durch die Jury der GESELLSCHAFT FÜR DEUTSCHE SPRACHE selbst zu einem Gegenstand der Berichterstattung wird, kann man es 66 mal in 32 Dokumenten finden. Im darauf folgenden Monat wird es ebenfalls im Zusammenhang mit der Auswahl zum Unwort noch 45 mal in 22

Dokumenten aufgeführt. In den folgenden fünf Monaten ist das Erscheinen mit 63 Token in 31 Zeitungsartikeln marginal. Ab August 2002 wird das Wort jedoch wieder häufiger verwendet und tritt besonders stark in Erscheinung, wenn mutmaßliche Terroristen verhaftet werden. Im Dezember 2002 ist der Gebrauch mit 174 Nennungen in 69 Dokumenten noch einmal besonders stark frequentiert, da hier eine Todesliste des amerikanischen Geheimdienstes veröffentlicht wird, auf der Namen von Terroristen verzeichnet sind.

Aus den Beobachtungen zu diesem Experiment lässt sich erschließen, dass es gut gelingt, die Verwendung eines Wortes mit Hilfe von Zeitungskorpora auf einer Zeitachse darzustellen. Allerdings ist in diesem konkreten Fall der Erklärungswert einer solchen Darstellung relativ gering. Es lässt sich mit Blick auf die Häufigkeitsverteilung zwar vermuten, dass die Ächtung des Wortes durch die GESELLSCHAFT FÜR DEUTSCHE SPRACHE in den ersten vier Monaten noch zu einer selteneren Verwendung geführt haben könnte und anschließend wieder unbefangener gebraucht wird, wirklich belegen lässt sich eine solche Aussage aber weder im Hinblick auf die Häufigkeitsverteilung noch durch eine inhaltliche Analyse der Kontexte, in denen das Wort auftritt.

Für die korpusbasierte Lexikografie bedeutet dieses Experiment, dass neue Angaben in Wörterbüchern zur Diachronie einzelner Lexeme nur möglich werden, wenn diese eine bestimmte Häufigkeitsklasse überschreiten. Selbst wenn man Korpora noch größer gestaltet als sie gegenwärtig sind, werden Ergebnisse zu seltenen Wörtern bereits durch einzelne Quellen und Belege so stark verfälscht, dass Aussagen zu ihrer Verwendung kaum zu treffen sind. Schließlich muss auch beachtet werden, dass Phänomene des Sprachwandels sich in der Regel über einen längeren Zeitraum erstrecken als in dem hier beschriebenen Experiment. Ein elektronisches Monitorkorpus, dessen Textbasis homogen mehrere Dekaden abdeckt, gibt es für die deutsche Sprache bislang nicht.

4.2.3. Neologismen

Als Neologismen werden lexikalische Einheiten bezeichnet, die entweder eine ganz neue Gesamtheit aus Form und Bedeutung bilden (z.B. *Handy* für mobiles Telefon) oder einer etablierten lexikalischen Einheit eine neue Bedeutung verleihen (z.B. *Wende* für den politischen Wandel nach 1989).[157] Traditionell findet ein Lexikograf Neuprägungen und Umdeutungen, indem er aus Texten solche sprachlichen Einheiten exzerpiert, die er subjektiv für Neologismen hält.[158] Auch wenn mehrere Lexikografen zusammenarbeiten, werden auf

[157] Vgl. KINNE 1998, S. 85.
[158] Vgl. TEUBERT 1998, S. 130.

diese Weise viele Neuschöpfungen übersehen. Daher versucht man auf der empirischen Basis von Textkorpora Kandidaten zu finden, die als Neologismen in Frage kommen.[159]

TELLENBACH beschreibt, wie eine Projektgruppe am IDS versucht, Neologismen der 90er Jahre mit Hilfe eines Referenzkorpus zu identifizieren, indem eine Liste mit Stichwort-kandidaten an einem Korpus mit gegenwartssprachlichen Texten aus der Zeit vor 1991 abgeglichen wird.[160] Lexeme, die bereits im älteren Vergleichskorpus mehrfach belegt sind, können in der Regel aus der Liste der Neuschöpfungen gelöscht werden. Probleme gibt es z.B. bei Neologismen, die (häufig aus dem Englischen) entlehnt oder übertragen werden, da diese einen Integrationsprozess durchlaufen, der neben mehreren Schreibweisen und Abkürzungsvarianten auch verschiedene Varianten der Kompositabildung zeigen kann, wie es TELLENBACH z.B. für das Wort *Anchorman* identifiziert, das auch als *Anchormann* und sogar als *Ankermann* in den Belegstellen auftritt.[161] An diesem Beispiel wird deutlich, dass eine solche Methode stark an der Ausdrucksseite der Lexeme orientiert ist und daher nur „echte" Neologismen erkennen kann.

TEUBERT beschreibt ein Monitoring-Verfahren, das darauf ausgerichtet ist, Veränderungen auch auf der Bedeutungsseite von Wörtern festzustellen, um Neologismen zu finden, die einen bereits vorhandenen Ausdruck lediglich umdeuten.[162] Dabei wird das Korpus zunächst in Phasen zerlegt, um die absolute und relative Häufigkeit jedes Types in einzelnen Zeitabschnitten zu ermitteln. Wenn sich die Häufigkeit im Übergang zu einer angrenzenden Phase in statistisch relevantem Maße ändert, ist ein Kandidat für einen Bedeutungswandel gefunden. In einem zweiten Schritt wird der Kontext des Types in den jeweiligen Phasen analysiert, indem geprüft wird, ob sich signifikante Veränderungen in den Häufigkeiten der rechten und linken Kollokate vollzogen haben, bzw. ob neue Kollokate hinzutreten. Ist dies der Fall, hat der Lexikograf genügend Anhaltspunkte zu der Annahme, dass hier ein Neologismus vorliegen könnte, den er manuell zu prüfen hätte, um zu entscheiden, ob man ihn in ein Wörterbuch aufnehmen muss.

Auch beim anschließenden Verfassen des Wörterbuchartikels können Korpora wiederum sehr hilfreich sein, wenn es darum geht, sich für eine Grammatikangabe zu entscheiden, falls sich beispielsweise für einen Neologismus Genus-Varianten finden oder nicht klar ist, ob ein Anglizismus bei der Flexion im tatsächlichen Sprachgebrauch durch ein Genitiv *-s* markiert wird.[163]

---

[159] Vgl. TEUBERT 1998, S. 130.
[160] Vgl. TELLENBACH 2001, S. 106ff. (Hervorhebungen im Original).
[161] Vgl. TELLENBACH 2001, S. 107.
[162] Vgl. TEUBERT 1998, S. 159.
[163] Vgl. TELLENBACH 2001, S. 112.

Probleme kann es bei der Lemmatisierung geben, wenn man Neologismen in das lexikografische Projekt aufnehmen will, für die sich noch keine Schreibnorm etabliert hat.[164] Die Lemmazeichengestaltangabe wird noch einmal zusätzlich erschwert, wenn Variantenschreibung zugelassen ist, ein Beispiel hierfür ist der Bindestrich bei Kompositabildung, der häufig bei Anglizismen wie *Shareholder-Value* auftritt, aber auch wie in *Golden Goal* weggelassen wird.[165]

Es ist festzuhalten, dass korpusbasierte Verfahren offenkundig ein Problem lösen, das sich mit jeder Neuauflage eines Wörterbuchs stellt, wenn ermittelt werden muss, welche Wörter in die Lemmaliste neu aufgenommen werden. Für Wörterbuchbenutzer könnte es weiterhin interessant sein, im Lexikon eine Belegangabe vorzufinden, die darüber informiert, für wann die früheste Verwendung eines Wortes bekannt ist.[166]

4.2.4. Fachsprache

HEYER/QUASTHOFF/WOLFF beschäftigen sich mit der Frage, welche Möglichkeiten es gibt, aus Fachtexten automatisch Fachtermini herauszufiltern, um sie systematisiert in elektronischen Terminologie-Wörterbüchern zu kompilieren.[167] Sie identifizieren zwei Gruppen von Ansätzen zur automatischen Gewinnung von Fachbegriffen aus Textkorpora:

a) Symbolisch/syntaktische Verfahren mittels ‚Pattern-Matching' (Mustererkennung);
b) Statistische Verfahren.[168]

Die symbolischen Verfahren stützen sich auf die Beobachtung, dass es morphologische Komponenten gibt, die für Fachterminologie typisch bzw. statistisch signifikant häufig sind wie z.B. *anti-, mega-, mikro-* als Präfixe oder *-ial, -ik, -tät* als Suffixe. Mit Hilfe solcher Morphemlisten lassen sich Suchabfragen gestalten, mit denen man aus Textkorpora Fachtermini filtern kann.[169] Symbolische Verfahren kommen mit relativ kleinen Korpora aus. Die Qualität ihrer Ergebnisse ist von der Vollständigkeit der Morphemlisten abhängig, die mühevoll in Handarbeit erstellt werden müssen.[170]

Im Kontrast hierzu stehen die statistischen Verfahrensweisen, die sich darauf stützen, dass Fachtermini in Fachliteratur häufiger vorkommen als in allgemeinsprachigen Korpora, sodass die Möglichkeit besteht, mit Hilfe eines Abgleichs von Häufigkeiten Fachtermini

[164] Vgl. TELLENBACH 2001, S. 108.
[165] Vgl. TELLENBACH 2001, S. 109.
[166] Vgl. TELLENBACH 2001, S. 110.
[167] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 43.
[168] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 44.
[169] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 44.
[170] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 48.

aufzuspüren, wobei es allerdings noch nicht genügt, Frequenzlisten der häufigsten Wörter gegenüberzustellen. Es wird vielmehr versucht, nach einem festgelegten Schlüssel jene Wortformen aufzuspüren, die relativ zu einem standardsprachlichen Referenzkorpus den Schwellenwert einer bestimmten Häufigkeitsklasse erreichen.[171] Da Fachtermini meist Substantive sind, lassen sich die statistisch ermittelten Rohdaten durch Filter optimieren, die darauf ausgerichtet sind, andere Wortarten zu erkennen und aus dem Datenbestand zu nehmen.[172] Die Erkennung anderer Wortarten könnte beispielsweise wiederum durch ein morphologisches Pattern-Matching erfolgen, da kaum davon auszugehen ist, dass Fachtexte in der Regel linguistisch annotiert vorliegen. Statistische Verfahren benötigen relativ große Korpora und die Qualität der Ergebnisse ist von der Wahl des Vergleichskorpus und des richtigen Schwellenwertes abhängig.[173]

In einem Verfahren, das zu Fachbegriffen automatisch Definitionen in Fachtexten aufspüren soll, verwenden HEYER/QUASTHOFF/WOLFF Kollokationen, indem sie die häufigsten linken und rechten Wortnachbarn identifizieren und diese Kollokate wiederum nach ihren häufigsten Nachbarn befragen, sodass ein Bedeutungsnetz mit den relevantesten Bedeutungsbeziehungen entsteht. Über ein Formular kann dieses Bedeutungsnetz anschließend von einem Benutzer manuell abgefragt werden, um genau jene Sätze zu finden, die einer Definition des Fachbegriffes am nächsten kommen, für den er sich interessiert.[174] Dieses Verfahren bietet gegenüber herkömmlichen Fachwörterbüchern den Vorteil, dass die Fachtexte selbst Auskunft über die Verwendungsweisen der Terminologie liefern.

Zusammenfassend kann festgestellt werden, dass Fachwörterbücher durch die hier beschriebenen Verfahrensweisen sicherlich ergänzt und bereichert werden können. Ob es zweckdienlich ist, Fachbegriffe ausschließlich in dieser Weise zu ermitteln und zu definieren, darf jedoch bezweifelt werden, da Fachwörterbücher auch der Explikation von Phänomenen dienen, die in Fachpublikationen zu kompliziert beschrieben sind.

4.2.5. Wörterbuchwissen aus Wörterbuchtext

Da Wörterbücher, die als ‚Machine Readable Dictionary' (MRD) z.B. auf CD-ROM vorliegen, selbst Korpora bilden können und in ihnen bereits Wissen über Wörter systematisiert ist, liegt es nahe, einen Computer auf dieses Wissen zugreifen zu lassen, denn MRDs stellen zur Zeit vermutlich das größte verfügbare lexikalische Wissensrepertoire

[171] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 45.
[172] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 46.
[173] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 48.
[174] Vgl. HEYER/QUASTHOFF/WOLFF 2002, S. 46ff.

bereit.[175] Allerdings sind MRDs nicht für den Computer, sondern für den Menschen konzipiert. Vieles, was diesem bei der Benutzung von MRDs durch Transfer- und Interpretationsleistungen zu erschließen möglich ist, kann ein Computersystem nicht nachvollziehen.[176] MRDs sind auch trotz ihres Umfangs ganz und gar nicht vollständig, sondern liefern nur Ausschnitte der Sprache, die einem Computer als willkürlich gewählt gelten müssen, da „schwierige" Wörter umfassender beschrieben werden als Wörter, deren Gebrauch den Menschen einfach erscheint.[177] Wegen solcher Probleme kann keine zuverlässige Schnittstelle zwischen dem Datenbestand und Regeln zur Sprachverarbeitung programmiert werden. Weiterhin ist unklar, wie die Syntax- und Semantiktheorien, die durch ein Lexikon repräsentiert werden, in ein automatisches Sprachverarbeitungssystem überführt werden könnten. Schließlich ist es bedenklich, ein bestehendes Lexikon als Basis für die Auflage eines neuen Wörterbuchs zu verwenden, da sich der vollzogene Sprachwandel nicht ausdrücken kann.

Trotz dieser Schwierigkeiten kann es durchaus sinnvoll sein, das in MRDs enthaltene Wissen einzusammeln, um es beispielsweise in lexikalischen Datenbanken[178] wiederzuverwerten. Der erste Schritt zur Extraktion des Wissens aus dem Wörterbuchtext führt über die Konvertierung des Satzbandes in ein Datenformat, das sich zur Auswertung eignet (Reformatierung).[179] Spezielle Wörterbuchparser, die individuell für die jeweiligen MRDs entwickelt werden müssen, folgen dem Anspruch, im Satzband beinhaltete Informationen so zu transformieren, dass sie für den Lexikografen auswertbar editiert werden.[180] Der anschließenden Extraktion des sprachlichen Wissens muss eine möglichst präzise Analyse der Mikrostruktur[181] des MRDs vorausgehen. Verhältnismäßig einfach ist es dann beispielsweise die Extraktion pragmatischer Angaben vorzunehmen, die in den Wörterbuchartikeln oft durch entsprechende Abkürzungen markiert und daher leicht zu identifizieren sind.

Bei ELSEN/HARTMANN findet sich ein Ansatz, aus digital vorliegenden Satzbändern von Printwörterbüchern aus der Typografie und der unmittelbaren Umgebung eines Elements der Mikrostruktur lexikalisches Wissen zu extrahieren.[182] Sie stellen beispielsweise fest, dass das erste Objekt in Kursivschrift hinter einem Lemma oder einem Lesartbezeichner immer eine Bedeutungsangabe ist.

---

[175] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 7.
[176] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 8.
[177] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 8.
[178] Vgl. S. 66
[179] Vgl. HEYN 1992, S. 3.
[180] Vgl. HEYN 1992, S. 2ff.
[181] Vgl. S. 57
[182] Vgl. ELSEN/HARTMANN 1993, S. 173ff.

WEBER versucht dagegen, aus Definitionstexten, die sich in der Mikrostruktur von Lexikonartikeln befinden, Synonymangaben für Lemmata zu extrahieren, indem er die Definitionstexte nach ihren Quantitäten klassifiziert.[183] Beispielsweise enthalten Einwortdefinitionen häufig Wörter, die als bedeutungsgleich gelten können (*Ohrfeige ~ Backpfeife*), während Dreiwortdefinitionen kumulativ Synonymie explizieren (*albern ~ einfältig, töricht, kindisch*).

Zusammenfassend ist festzuhalten, dass die Übernahme von Angaben aus Lexikoneinträgen, die in MRDs zur Verfügung stehen, nicht ohne beträchtlichen manuellen Aufwand möglich ist, wenngleich es sinnvoll sein kann, bereits bestehende Lexika zu konsultieren, um nicht auf der Basis von Textbelegen den gesamten deutschen Wortschatz neu beschreiben zu müssen.[184]

[183] Vgl. WEBER 1993, S. 145ff.
[184] Vgl. LANGER 1995, S. 52.

5. Erschließung von Bedeutung

In der Korpuslinguistik sind die größten Schwierigkeiten und Chancen mit der Inhaltsseite von Sprachzeichen verbunden. Einerseits hat man enorme Probleme, Korpora semantisch zu annotieren, andererseits lassen sich gerade durch korpusbasierte Verfahren Aussagen über die Grundbedeutungen der Wörter treffen.

In der Lexikografie sind die Schwierigkeiten und Chancen ganz ähnlich verteilt. Einerseits hat man enorme Probleme, Bedeutung kohärent und adäquat in Wörterbüchern zu beschreiben, andererseits lässt sich nirgends so schnell etwas über die Bedeutung von Wörtern erfahren wie in Lexika.

Die folgenden Beschreibungen gehen von der Hypothese aus, dass sich Schwierigkeiten und Chancen der korpusbasierten Lexikografie genau an jenem Punkt kreuzen, wo es sowohl dem kompetenten Sprecher als auch dem computerlinguistischen Verfahren gelingen sollte, Bedeutung zu disambiguieren.

5.1. Zur Struktur des Lexikonartikels

Linguisten ist frühzeitig aufgefallen, dass der Artikelteil konventioneller Lexika nicht ideal strukturiert ist, um die Bedeutungsseite des Wortschatzes widerzuspiegeln. So haben bereits KATZ/FODOR versucht, eine semantische Theorie zu entwerfen, in der Lexikoneinträge derart gestaltet sind, dass alle semantischen Eigenschaften und Relationen von Wörtern formal repräsentiert werden, um die Leistung eines kompetenten Sprechers bei der Disambiguierung von Mehrdeutigkeit abzubilden.[185]

Im Artikelteil des gedruckten Wörterbuchs erscheinen Angaben, die sich auf die Form eines Wortes beziehen, im Kopf des jeweiligen Eintrags, wenn zunächst Aussagen zu Orthografie und Morphologie getroffen werden. Aussagen zu der Inhaltsseite des Sprachzeichens erscheinen anschließend, wenn die jeweiligen Lesarten[186] aufgeführt werden. Diese Strukturierung des Lexikonartikels wird durch die Zweidimensionalität des Druckraums erzwungen und suggeriert ungewollt, dass das Lemmazeichen der zentrale Gegenstand ist, über den etwas ausgesagt werden muss.

HAß-ZUMKEHR fragt, ob das digitale Wörterbuch nicht wesentlich geeigneter als das gedruckte sei, eine konsequente Zuordnung von formbezogener und inhaltsbezogener Information zu leisten, da man davon ausgehen müsse, dass der Computer, ähnlich wie der kompetente Sprecher, den Lesarten eines Wortes z.B. durch assoziierten Kontext jeweils mehr

[185] Vgl. KATZ/FODOR 1970.
[186] Vgl. S. 50.

oder weniger Identität zuschreiben könne.[187] Ein wesentliches Argument für die an Lesarten orientierte Strukturierung eines Wörterbuchs sieht HAß-ZUMKEHR darin, dass paradigmatische und syntagmatische Beziehungen ausschließlich zwischen Lesarten und nicht zwischen Lemmazeichen existieren.[188]

Es wird hier deutlich, dass man sich von Korpora die Chance erhoffen kann, dass die Lesarten der Lexeme aufgespürt werden, um sie in einem digitalen Wörterbuch in einer Struktur zu organisieren, die näher an der Sprachkompetenz des Menschen liegt, indem sie mehr Auskunft über die Bedeutungsbeziehungen von Wörtern gibt als traditionelle Wörterbuchartikel.

5.2. Lesarten

HAß-ZUMKEHR beschreibt das zentrale sprachtheoretische Problem, mit dem sie wiederholt konfrontiert wird, durch die Frage: „*Was ist der zentrale Gegenstand lexematischer Information - die gesamte Inhaltsseite des Lemmazeichens oder die jeweilige Einzelbedeutung?*“[189] Die jeweiligen Einzelbedeutungen der Wörter lassen sich auch als Lesarten bezeichnen, die von ENGELBERG/LEMNITZER definiert werden als „*von den LexikograhInnen zusammengefasste Mengen von Verwendungsweisen oder Verwendungskontexten, die untereinander als hinreichend ähnlich und als hinreichend verschieden von allen anderen Gebrauchsweisen empfunden werden.*“ [190]

Lesartendisambiguierung ist eine wichtige Voraussetzung für Sprachkompetenz, da gerade häufig verwendete Wörter mehrdeutig sind und dem Menschen Entscheidungen abverlangen, welche Einzelbedeutungen adäquat auszuwählen sind. Auch für den Computer ist Lesartendisambiguierung der entscheidende Prozess beim Entschlüsseln von Sprache.[191]

Für die traditionelle Lexikografie lässt sich feststellen, dass Lexikografen bei der Auswahl von Lesarten intuitiv vorgehen müssen und Wörterbücher selten in der Beschreibung von Lesarten übereinstimmen. Die korpusbasierte Lexikografie kann sich dagegen der statistischen Eigenschaften von Texten bedienen, indem sie die Korpusdaten in Klassen gruppiert, aus denen dann Lesarten geformt werden. ENGELBERG/LEMNITZER berichten, dass zur Zeit an Verfahren gearbeitet wird, die eine lexikalisch-semantische Disambiguierung erlauben, indem Ähnlichkeiten in den Belegstellen von Wörtern so analysiert und

[187] Vgl. HAß-ZUMKEHR 2001, S. 113.
[188] Vgl. HAß-ZUMKEHR 2001, S. 114.
[189] Vgl. HAß-ZUMKEHR 2001, S. 113.
[190] ENGELBERG/LEMNITZER 2001, S. 208.
[191] Vgl. KUNZE/WAGNER 2001, S. 239.

zusammengestellt werden, dass aus den Vorkommenstexten Entscheidungshilfen für oder gegen die Einzelbedeutungen von Lexemen erschlossen werden können.[192]

Hier wird deutlich, dass sich aus korpusbasierten Verfahren zahlreiche Chancen für die Sprachtechnologie entwickeln, wenn es ihnen gelingt, Lesarten zu disambiguieren. Auch konventionelle Wörterbücher können von korpusbasierten Verfahren der Lesartengewinnung profitieren. So ist es in der britischen Lernerlexikografie üblich geworden, Lesarten nach der Häufigkeit ihres Auftretens anzugeben.[193]

5.3. Selektionsbeschränkungen

Lesarten können mit Hilfe von Selektionsbeschränkungen disambiguiert werden. Als Selektionsbeschränkungen gelten semantische Auswahlbedingungen, die ein Prädikat an seine Argumente stellt.[194] So verlangt beispielsweise das Verb *fürchten* in der Regel nach einem belebten Agens, während das, was gefürchtet wird, beliebiger Natur sein kann. Bei dem Verb *blättern* wird dagegen eher ein unbelebtes Patiens (z.B. *Buch*) auftreten müssen, während das Agens sowohl belebt (z.B. *Kind*) als auch unbelebt (z.B. *Wind*) sein kann. Die Beschränkungen bei der Selektion von Agens und Patiens leisten einen wichtigen Beitrag zur semantischen Disambiguierung. Der Satz

Das Haus fürchtet das Kind.

entzieht sich aus Sicht von Morphologie und Syntax der Beurteilung von Nominativ und Akkusativ. Nur durch die semantischen Selektionsbeschränkungen von *fürchten*, das nach einem belebten Agens verlangt, kann disambiguiert werden, wie die Satzteile zu bestimmen sind.

Selektionsbeschränkungen sind wichtige Informationsquellen für die Disambiguierung in der maschinellen Sprachverarbeitung.[195] Man versucht hier nicht bloß Selektionsbeschränkungen zu finden, sondern Selektionspräferenzen durch einen Wert zu charakterisieren, der durch statistische Analyse großer Korpora ermittelt wird.[196] Selektionsbeschränkungen können aber auch für menschliche Benutzer von Lexika zweckmäßig sein, wenn es beispielsweise darum geht, den Gebrauch eines Fremdwortes zu klären. Selektionsbeschränkungen befinden sich in den meisten Lexikoneinträgen bisher nur implizit in den Kommentaren[197]. Korpuslinguistik könnte hier eine neue Angabeklasse etablieren, wobei aber davon auszugehen ist, dass sich

[192] Vgl. ENGELBERG/LEMNITZER 2001, S. 208.
[193] Vgl. ENGELBERG/LEMNITZER 2001, S. 208.
[194] Vgl. KUNZE/WAGNER 2001, S. 241.
[195] Vgl. KUNZE/WAGNER 2001, S. 241.
[196] Vgl. KUNZE/WAGNER 2001, S. 242.
[197] Vgl. S. 57

Selektionsbeschränkungen nicht durch ein kleines Inventar von Merkmalen wie *belebt* oder *unbelebt* beschreiben lassen, sondern aufwendig manuell erfasst werden müssen, da prinzipiell jede semantische Eigenschaft bei der Definition der Selektionsbeschränkungen eines Wortes eine Rolle spielen kann.[198]

5.4. Bedeutungsressourcen

Innerhalb der maschinellen Sprachverarbeitung stehen heute lexikalische Wissensbasen im Zentrum der Aufmerksamkeit, weil die darin enthaltenen Strukturierungen von Wortbedeutungen eine unabdingbare Voraussetzung für zahlreiche Anwendungsgebiete der Computerlinguistik darstellen.[199]

Die lexikalisch-semantische Ressource GERMANET[200], die an der Universität Tübingen erstellt wird, bedient sich der Organisationsprinzipien, die in Princeton mit WORDNET entwickelt worden sind, insofern sie das Datenbankmodell und die Struktur betreffen.[201] WORDNET bildet Lexeme in ihren vernetzten Bedeutungsbeziehungen ab, um für computerlinguistische Verfahren ein umfassendes Referenzwissen bereitzustellen, das durch Abgleich mit Textwörtern der Bedeutungsdisambiguierung dient.[202] GERMANET ist keine Übersetzung von WORDNET, sondern aus deutschen lexikografischen Quellen mit Berücksichtigung von Korpusfrequenzen erarbeitet worden.[203] KUNZE/WAGNER beschreiben die Unterschiede zwischen GERMANET und WORDNET im Wesentlichen wie folgt:

- GERMANET orientiert sich im Gegensatz zu WORDNET an linguistischen und nicht an psychologischen Strukturprinzipien;
- in GERMANET wird von artifiziellen Konzepten systematisch Gebrauch gemacht, um die Konzepthierarchie ausgewogener zu gestalten;
- in GERMANET werden im Gegensatz zu WORDNET Partikelverben kodiert;
- Kausationsbeziehungen werden in GERMANET nicht nur für Verben, sondern Wortart übergreifend dargestellt;
- Adjektive werden taxonomisch strukturiert und nicht wie in WORDNET in einem assoziativen Verbund.[204]

[198] Vgl. KUNZE/WAGNER 2001, S. 241ff.
[199] KUNZE/WAGNER 2001, S. 229.
[200] Vgl. Website von GERMANET http://www.sfs.nphil.uni-tuebingen.de/lsd/
[201] Vgl. FELLBAUM 1998
[202] Vgl. KUNZE/WAGNER 2001, S. 229.
[203] Vgl. KUNZE/WAGNER 2001, S. 230.
[204] Vgl. KUNZE/WAGNER 2001, S. 235.

In GERMANET ist der deutsche Grundwortschatz auf konzeptueller Ebene so modelliert, dass Nomen, Verben und Adjektive in ihren elementaren Bedeutungsbeziehungen dargestellt werden.[205] Das zentrale Konzept der lexikalischen Kodierung von GERMANET liegt in den „*so genannten synsets, die als abstrakte Bedeutungseinheiten zu gegebenen Konzepten eine Synonymenmenge bereitstellen.*“[206] Die wichtigste semantische Relation, durch die Synsets und Wortbedeutungen (einzelne Synonyme aus den Synsets) gegliedert werden, ist die hierarchische Strukturierung der Konzepte aller drei Wortarten durch Hyponymie-Beziehungen.[207] Einige weitere semantische Beziehungen, die dargestellt werden, sind Meronymie (Teil-Ganzes-Beziehung) für Nomen, Kausationsbeziehungen (*töten-sterben*, *öffnen-offen*) vor allem für Verben und semantische Derivationsrelationen, die Kategorien übergreifend für denominale Adjektive, deverbale Nominalisierungen und deadjektivische Nominalisierungen gelten.[208]

Name	Symbol	Valid Class			Reverse Pointer		Lex
		N	A	V	Name	Symbol	
Antonym	!	y	y	y	Antonym	!	y
Hyperonym	@	y	y	y	Hyponym	~	n
Hyponym	~	y	y	y	Hyperonym	@	n
Meronym	%	y	n	n	Holonym	#	n
Holonym	#	y	n	n	Meronym	%	n

Tabelle 5. ‚Semantic Pointers' in GERMANET (Auswahl)

Datenquelle: Website von GERMANET

Tabelle 5 zeigt, wie einige semantische Relationen in den Konzepten notiert werden. Zunächst ist die Art einer Relation benannt, der ein Symbol zugeordnet wird. In den folgenden drei Spalten ist angegeben, für welche Wortart die Relation gültig ist. Unter der Spalte “Reverse Pointer” wird die semantische Relation in umgekehrter Richtung verfolgt. Die Spalte „Lex“ gibt an, ob sich die semantische Relation nur auf ein Element oder auf das ganze Konzept bezieht.

Konzepte, die unterschiedlichen Hierarchieebenen zuzuordnen sind, werden in tieferen Schichten über Kreuz klassifiziert, wodurch Muster von Polysemie deutlich werden, die in einer bloßen Baumstruktur nicht dargestellt werden könnten.[209] Wenn Sprache nicht genügend Kohyponyme bereitstellt, um eine untere Schicht der Hierarchie angemessen in Relation zu

[205] Vgl. KUNZE/WAGNER 2001, S. 230.
[206] KUNZE/WAGNER 2001, S. 230.
[207] Vgl. KUNZE/WAGNER 2001, S. 231.
[208] Vgl. KUNZE/WAGNER 2001, S. 230.
[209] Vgl. KUNZE/WAGNER 2001, S. 232.

setzen, werden artifizielle Konzepte eingeführt, die als Knotenpunkte dienen.[210] In *Abbildung 5* werden zwei solcher Knoten gezeigt, denen es gelingt Ko-Hyponyme einzuordnen.

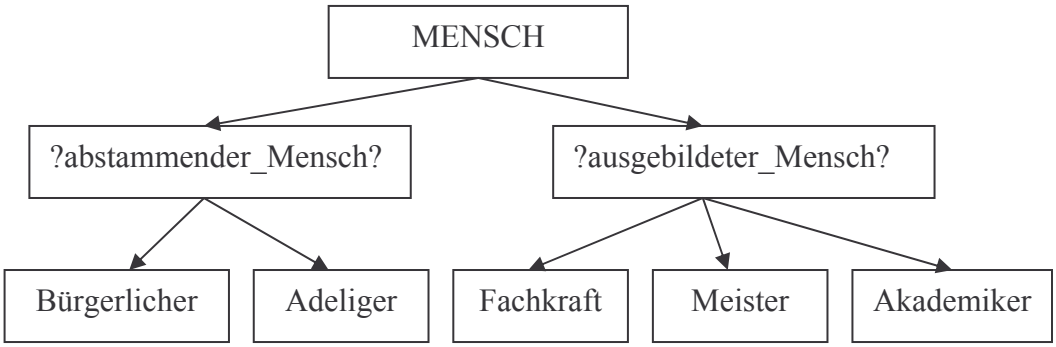


Abbildung 5. Konzept mit artifiziellen Knoten für ein Hauptwort
Quelle: Frei nach Website von GERMANET

Verbeinträge werden in GERMANET immer mit Subkategorisierungsrahmen versehen, deren Notation sich an den CELEX-Frames[211] orientiert. Die kodierten Rahmen können bei der Lesartendisambiguierung helfen, indem sie Aufschluss über das syntaktische Komplementierungsverhalten der GERMANET-Prädikate leisten und bilden so einen Beitrag zur Syntax-Semantik-Schnittstelle.[212]

In GERMANET sind Synsets für ca. 27.000 Nomen, 9.000 Verben und 5.000 Adjektive in Datenbanken gespeichert, welche mit Frequenzlisten, die aus Korpora extrahiert worden sind, systematisch abgeglichen werden.[213] Die Ressource GERMANET kodiert nur morphologische Vollformen, beinhaltet lediglich sehr geläufige Mehrwortlexeme und markiert Eigennamen und Abkürzungen gesondert.

GERMANET ist in das Projekt EUROWORDNET[214] integriert, das die semantisch-lexikalischen Netze mehrerer europäischer Sprachen in einer Datenbank zusammenfasst und untereinander verbindet.

[210] Vgl. KUNZE/WAGNER 2001, S. 234.
[211] Vgl. S. 26
[212] Vgl. KUNZE/WAGNER 2001, S. 234.
[213] Vgl. KUNZE/WAGNER 2001, S. 231.
[214] Vgl. Website von EUROWORDNET http://www.illc.uva.nl/EuroWordNet/

6. Kompilierung

Der Begriff ‚Kompilierung' ist in der Literatur nicht klar umrissen, sodass hiermit sowohl die Gesamtheit aller nötigen Basisoperationen bezeichnet sein kann, die das sprachliche Wissen aus Texten in Wörterbücher überführen als auch das bloße Editieren von bereits strukturiert vorliegenden Erkenntnissen. Hier soll Kompilierung daher nur als Überschrift Verwendung finden, um die ganze Bandbreite an Überlegungen und Operationen darzustellen, die nötig sind, das extrahierte sprachliche Wissen in eine Wörterbuchstruktur zu überführen.
Im Mittelpunkt der Betrachtungen sollen dabei Konstruktion, Auswahl und Anordnung der Lexikoneinträge stehen, da alle lexikografischen Projekte diesbezüglich mit ähnlichen Fragen konfrontiert sind.[215]

6.1. Publikationsmodell

Obwohl sich die Lexikografie zunehmend mit den Möglichkeiten digitaler Wörterbücher auseinandersetzt, sind elektronische Produkte noch immer stark an Strukturen orientiert, die sich in gedruckten Medien herausgebildet haben.[216] MÜLLER/SCHMIDT versuchen ein lexikografisches Modell zu entwerfen, das sich die Erkenntnisse aus der Printproduktion zu Nutze macht, jedoch für digitale Repräsentation umgestaltet wird.[217] Sie fordern ein Publikationsmodell, das die Ebenen der Inhaltsstruktur deutlich von denen der Redaktion und des Benutzers trennt. Für die Inhaltsstrukturmodellierung wird dabei nicht Standardisierung im Sinne von Vereinfachung verlangt, sondern die konsistente Erfassung der Komplexität neuer Inhalte.[218] Um zu verdeutlichen, wie das gemeint ist, bietet sich die Unterscheidung zweier strategischer Varianten an, die STORRER für die Informationsmodellierung in Opposition stellt:

a) Die Modellierung des digitalen Wörterbuchs orientiert sich in der Makro- und Mikrostruktur an den Vorgaben, die in der Lexikografie für gedruckte Wörterbücher entwickelt worden sind. Die wichtigsten Bauteile sind dann die einzelnen Artikel.
b) Die Modellierung eines digitalen Wörterbuchs orientiert sich an dem Gegenstand, den es beschreibt. Die Makro- und Mikrostrukturen richten sich dann an den lexikalischen Einheiten aus. Als wichtigste Bauteile gelten hier Grapheme, Morpheme und Lexeme.[219]

[215] Vgl. KRISHNAMURTHY 1987, S. 62.
[216] Vgl. STORRER 2001, S. 54.
[217] Vgl. MÜLLER/SCHMIDT 2001, S. 49.
[218] Vgl. MÜLLER/SCHMIDT 2001, S. 50.
[219] Vgl. STORRER 2001, S. 60.

STORRER entscheidet sich für die zweite Variante, wenn sie fordert, dass sich die konzeptuelle Datenmodellierung digitaler Wörterbücher nicht an Strukturen von Printprodukten ausrichten sollte, sondern an linguistischen Einheiten.[220] Die Vorteile dieser Entscheidung liegen darin, dass erstens viele Eigenschaften lexikalischer Einheiten (z.B. Flexion, Derivation) nicht mehr zu jedem Lemma einzeln aufgeführt werden müssen, sondern bei Bedarf über Regeln generiert werden können. Zweitens ist ein linguistisch motiviertes Datenmodell flexibler in den Filtern der Abfrage. Drittens erfordert die linguistisch motivierte Informationsmodellierung ein höheres Maß an Formalisierung, was sich positiv auf der Benutzerseite auswirken kann, da hierdurch Mensch und Computer gleichberechtigt Möglichkeiten erhalten, auf das sprachliche Wissen zuzugreifen.[221]

6.2. Wörterbuchstrukturen

Um auszuloten, mit welchen Chancen und Perspektiven Lexikografie speziell durch die Arbeit mit Computersystemen bei der Kompilierung konfrontiert ist, wird hier wie von LEMBERG/SCHRÖDER/STORRER gefragt werden müssen, inwieweit „*die am Buch orientierten Traditionen der Wörterbuchgestaltung, z.B. die Makro-, Mikro-, Zugriffs- und Verweisstrukturen, im digitalen Medium noch zweckmäßig sind und inwieweit das neue Medium nicht nach anderen Gestaltungs- und Organisationsprinzipien verlangt.*“[222]

6.2.1. Makrostruktur

Im Artikelteil eines Wörterbuchs wird die Abfolge der Artikelstichwörter durch die Makrostruktur organisiert, wobei unabhängig von Fragen des Wortbegriffs oder des Lemma-Ansatzes allgemein orthografische Konventionen gelten, da Wörter in der Regel alphabetisch aufgelistet werden.[223] Begreift man die Architektur eines Wortschatzes als komplexe Ordnung, die auf einer Vielzahl von Strukturmerkmalen beruht, die sich oft auch kombinieren lassen und mit einer großen Zahl von Elementen, also Wörtern und ihren Verwendungsweisen operieren, dann wird verständlich, dass ein lexikologisches Organisationsprinzip nicht zwangsläufig zu einem bestimmten Ordnungsschema führen muss, sondern sich an einer Vielzahl von Interessen ausrichten kann.[224] So ist es beispielsweise denkbar, ein Wörterbuch

[220] Vgl. STORRER 2001, S. 61.
[221] Vgl. STORRER 2001, S. 61.
[222] LEMBERG/SCHRÖDER/STORRER 2001, S. 1.
[223] Vgl. SCHLAEFER 2002, S. 88.
[224] Vgl. GLONING/WELTER 2001, S. 118.

nicht nach der Schreibweise der Wortformen onomasiologisch zu strukturieren, sondern die Ordnung semasiologisch an Bedeutungsfeldern auszurichten.

Ein elektronisches Wörterbuch stellt Lexikologen und Lexikografen hinsichtlich der Makrostruktur vor neue Herausforderungen, da die Datenbasis derart mit Markierungen angereichert werden kann, dass eine Vielzahl von Interessen Berücksichtigung finden darf. Das Interesse menschlicher Benutzer an der Makrostruktur eines digitalen Wörterbuchs liegt vor allem in einer dynamischen Ordnung, die individuell wählbare Ansichten, Profile und Zugriffsmöglichkeiten gestattet. Die Anforderungen sprachverarbeitender Computersysteme machen es dagegen besonders wünschenswert, dass die Makrostruktur sämtliche Einträge des Lexikons miteinander in Verbindung setzt, um syntagmatische Beziehungen automatisch zu erkennen.[225] Aus Sicht der Informatiker ist die Makrostruktur wiederum eindeutig von dem jeweiligen Datenbankmodell abhängig und kann beispielsweise aus einer einfachen Liste, einer Tabelle oder einer Baumstruktur bestehen.[226]

6.2.2. Mikrostruktur

Der Artikelteil traditioneller Wörterbücher ist oft durch Stichwortreihe, Einzelartikel und Verweise dreigliedrig geebnet, wobei die einzelnen Artikel die größten selbständigen Informationseinheiten darstellen, die zumeist in einer systematisierten Mikrostruktur vorliegen, die für alle Artikel des jeweiligen Lemmazeichentyps gilt.[227] Die Mikrostruktur ergibt sich aus einer Klassifizierung elementarer Informationseinheiten.[228]

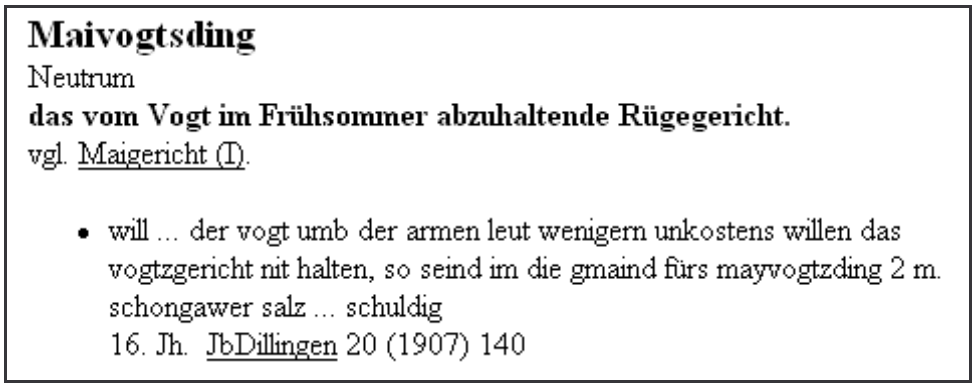
In der Mikrostruktur gibt es Angaben, die Fakten bereitstellen und Kommentare, die Gewichtungen, Interpretationen und Verständnishilfen bieten.[229] Kommentare setzen also mindestens eine Angabe voraus, auf die sie sich beziehen. Sie können Zusatzinformationen für grundsätzlich jede Angabe bieten, aber keine neuen Fakten einbringen.[230]

Während in gedruckten Wörterbüchern Angaben in funktionalen Textsegmenten der Mikrostruktur vorliegen, kann man in digitalen Wörterbüchern zusätzlich tabellarische Elemente und autonomen Text als Angabetypen ausmachen.[231] Als autonomen Text bezeichnet man jene Angaben, die dem Benutzer in identischer Form präsentiert werden wie z.B. Beschreibungen komplexer Pragmatik oder Semantik. Tabellarische Elemente dienen

[225] Vgl. SCHNEIDER 1999, S. 150.
[226] Vgl. GIBBON 2001, S. 397.
[227] Vgl. SCHLAEFER 2002, S. 84ff.
[228] Vgl. HAß-ZUMKEHR 2001, S. 108.
[229] Vgl. HAß-ZUMKEHR 2001, S. 107.
[230] Vgl. HAß-ZUMKEHR 2001, S. 108.
[231] Vgl. HAß-ZUMKEHR 2001, S. 108.

dagegen zur kontextfreien Klassifizierung beispielsweise von Flexionsparadigmen oder Varietätenzugehörigkeit.

Bei einem gedruckten Wörterbuch wird die Modellierung der Mikrostruktur an menschlichen Adressaten ausgerichtet. Dagegen ist bei digitalen Wörterbüchern die Strukturierung der Angaben auf die Belange der Informatik abgestimmt und wird durch die Präsentation in einer Front-End-Lösung dem menschlichen Benutzer nur indirekt zugänglich.[232]



Maivogtsding
Neutrum
das vom Vogt im Frühsommer abzuhaltende Rügegericht.
vgl. Maigericht (I).

- will ... der vogt umb der armen leut wenigern unkostens willen das vogtzgericht nit halten, so seind im die gmaind fürs mayvogtzding 2 m. schongawer salz ... schuldig
  16. Jh. JbDillingen 20 (1907) 140

Abbildung 6. Artikel des Onlinelexikons DEUTSCHES RECHTSWÖRTERBUCH
Quelle: Website von DEUTSCHESRECHTSWÖRTERBUCH

In *Abbildung 6* ist ein Artikel des Onlinelexikons DEUTSCHES RECHTSWÖRTERBUCH[233] dargestellt. Unter dem Lemma *Maivogtsding* bietet der Artikel eine Genus-Angabe, eine Bedeutungsangabe, einen Verweis und eine Belegstellenangabe. Hier zeigt sich zunächst, dass Konventionen des gedruckten Wörterbuchs noch weitgehend nachempfunden sind. Dass hinter dieser Präsentationsansicht eine Mikrostruktur liegt, die mit Printprodukten nur noch wenig gemeinsam hat, wird deutlich, wenn man sich den Quellcode dieses Eintrags betrachtet:

```
</UL><HR><A NAME="MAIVOGTDING" CLASS=STW>Maivogtsding</A><BR>
<SPAN CLASS=FN>Wortklasse:</SPAN>Neutrum
<BR><SPAN CLASS=FN>Erkl&auml;rung:</SPAN>
<SPAN CLASS=ERKL> das vom Vogt im Fr&uuml;hsommer abzuhaltende R&uuml;gegericht.
</SPAN><BR>
vgl.<A HREF="M12.htm#MAIGERICHT-1.0">Maigericht (I)</A>.<BR>
<UL CLASS=BEL><LI><SPAN CLASS=FN>Belegtext:</SPAN>
will ... der vogt umb der armen leut wenigern unkostens willen das vogtzgericht nit halten, so seind im
die gmaind f&uuml;rs <SPAN CLASS=BSTW>mayvogtzding</SPAN> 2 m. schongawer salz ...
schuldig<BR>
<SPAN CLASS=FN>Datierung:</SPAN> 16. Jh.
```

[232] Vgl. HAß-ZUMKEHR 2001, S. 104.
[233] Vgl. Website DEUTSCHES RECHTSWÖRTERBUCH http://www.rzuser.uni-heidelberg.de/~cd2/drw/index.htm

<SPAN CLASS=FN>***Fundstelle:***</SPAN> <A HREF="../s/I-kg.htm#JbDillingen">JbDillingen</A>

Hier zeigt sich, dass der Wörterbuchartikel nicht dreigliedrig geebnet ist, sondern in den Feldern *Wortklasse, Erklärung, Vergleich, Belegtext, Datierung* und *Fundstelle* gleichrangig strukturiert vorliegt.

Wenn ein digitales Wörterbuch nicht für Menschen, sondern für Computer gestaltet wird, entfernt sich die Struktur der einzelnen Artikel noch weiter von den Konventionen traditioneller Lexikografie. In *Tabelle 6* wird die Mikrostruktur eines Lexikons gezeigt, das aus verrauschten Textkorpora erstellt worden ist.[234] Eine derartige Datenpräsentation ist für Menschen wenig hilfreich, aber einer Spracherkennungssoftware[235] gestattet sie zu entscheiden, wie Wörter zu disambiguieren sind, die beim Einscannen von Dokumenten schlecht aufgelöst wurden. Dies gelingt, indem einem Lexem die hier dargestellten Wahrscheinlichkeitswerte, Frequenzen und Fundstellen zugeordnet werden, die anschließend ein Regelinventar durchlaufen, um das Wort zu disambiguieren.

	LEXEM/ÄHNLICHKEIT	HÄUFIGKEIT	TEXT(E)
STAMM	SENDEN	17	8,11,20,21,31,36,39,4,2, 46,49,54,56,67,7,6,78,87
VARIANTEN	zuzusenden / 0.4	10	4,16,19,32,40,61,83,83,9, 1,94,95
	übersenden / 0.4	7	7,9,23,45,48,77,100
	sendeni / 0.14	1	3
	scnden / 0.33	1	60
	sendungen / 0.33	1	74
	ende / 0.33	1	24
SUMME	7	38	38
RANG	9		

Tabelle 6. Mikrostruktur eines automatisch erstellten Lexikons

Datenquelle: SCHNEIDER 1999, S. 148

Abschließend ist anzumerken, dass die Gestaltung der Mikrostruktur eines korpusbasierten Wörterbuchs stark davon abhängt, welche Annotierungen an dem Korpus vorgenommen worden sind. Das Korpusdesign muss daher Anforderungen der Kompilierung antizipieren. Dies kann beispielsweise geschehen, indem die Mikrostruktur möglichst umfassend und differenziert ausgestaltet wird, um für diverse Projektmodule offen zu bleiben, auch für welche, die es noch nicht gibt.[236]

[234] Vgl. SCHNEIDER 1999.
[235] Gemeint ist hier eine Software zur ‚Opitcal Character Recognition' (OCR).
[236] Vgl. HAß-ZUMKEHR 2001, S. 104.

6.2.3. Angaben

Traditionelle Wörterbucheinträge ordnen üblicherweise den Lemmata verschiedene linguistische Angaben zu, die im Hinblick auf einen bestimmten Verwendungszweck des Lexikons sinnvoll sind, wobei es sich vor allem um phonetische, morphologische, syntaktische, semantische und pragmatische Angaben handelt.[237] Bei BIBER/CONRAD/REPPEN wird deutlich, dass korpusbasierte Lexikografie dieser Liste einige neue Klassen hinzuzufügen hat, nach denen die Autoren in folgender Weise fragen:

- Wie geläufig sind verschiedene Wörter in ihrer Verwendung?
- Wie gebräuchlich sind verschiedene Bedeutungen für ein bestimmtes Wort?
- Treten Wörter systematisch mit anderen Wörtern in Verbindung?
- Haben Wörter systematische Beziehungen zu bestimmten Textsorten oder Dialekten?[238]

Wenngleich herkömmliche Printwörterbücher zu solchen Fragen bereits sporadisch Auskunft erteilen, wäre es durch korpusbasierte Verfahren möglich, in Lexikoneinträgen durchgängig systematisiert diese Angabeklassen durch entsprechende Werte auszufüllen, die statistisch für jedes Wort und auch für Beziehungen zwischen Wörtern ermittelt werden könnten. Es wäre hier beispielsweise denkbar, nicht nur Wörter, sondern Wortverbindungen einer bestimmten Stilebene zuzuordnen, sodass dem Wörterbuchbenutzer Entscheidungshilfen geboten werden, wenn er nach einer angemessenen Formulierung sucht.

Da lexikalische Ressourcen zukünftig auch von Computern genutzt werden sollen, muss dort die Ausgestaltung von Angaben differenzierter und kohärenter gestaltet werden als in Wörterbüchern für Menschen. KRISHNAMURTHY beschreibt, wie das ‚Collins COBUILD English Language Dictionary' die Mikrostruktur ausgestaltet.[239] Hierbei wird deutlich, dass Lexikoneinträge eines korpusbasierten Wörterbuchs sich derart etikettieren lassen, dass nicht nur paradigmatische, sondern auch syntagmatische Beziehungen angegeben werden. So werden hier Lexeme beispielsweise mit dem Etikett VB für Verben, N für Hauptwörter, ADJ für Adjektive versehen. Weiterhin gibt es eine Notation für Beziehungen, in welche die Wörter untereinander treten, die bei KRISHNAMURTHY in etwa durch folgende Operatoren beschrieben werden:

+ = wird gefolgt von / hat in seiner Umgebung;
~ = wird gewöhnlich in einem bestimmten Tempus verwendet;
/ = ist lexikalisch verwirklicht als;
( ) = optionales Element von;

[237] Vgl. LENDERS/WILLÉE 1998, S. 105.
[238] Vgl. BIBER/CONRAD/REPPEN 1998, S. 21.
[239] Vgl. KRISHNAMURTHY 1987, S. 65ff.

<> = besteht aus / funktioniert als.[240]

Mit diesen Operatoren lassen sich Lexeme in Wörterbuchangaben so beschreiben, dass komplexe Aussagen darüber möglich werden, wie sie mit anderen Wörtern in Beziehung treten, was bei KRISHNAMURTHY in der folgenden Darstellung sichtbar wird:

N-AN//S = Hauptwort, das immer mit dem Artikel *a* oder *an* erscheint und niemals im Plural steht;
V+PREP/FOR = intransitives Verb, dem immer die Präposition *for* folgt;
PHR-VB<V-OD+ADV> = Phrasalverb, das aus einem transitiven Verb und einem Adverb besteht.[241]

Hier wird deutlich, dass sich korpusbasierte Lexikografie nicht allein auf neue Angabeklassen beschränken kann, die beispielsweise über Kollokate oder Frequenzen Auskunft geben. Wenn lexikalische Ressourcen erstellt werden, die komplexe Prozeduren durchlaufen sollen, müssen Angaben umfassend ausgestaltet werden.

6.2.4. Verweise

Auf der Ebene von Verweisen werden in dem Artikelteil eines Wörterbuchs vernetzende Strukturen zwischen Lemmata, Artikelinformationen oder externen Informationsquellen gebildet, die einerseits ergänzende Informationen und andererseits Möglichkeiten der Textkomprimierung bieten.[242] Das digitale Wörterbuch hat an diesen Stellen Verbindungsstrukturen (Hyperlinks), die den Anker und das Ziel idealerweise bidirektional verknüpfen.
Wenngleich es Ansätze gibt, diese Verbindungsstrukturen z.B. nach Art des Zieles, nach Art des Ankers oder nach Art der Funktion zu klassifizieren, hat sich noch keine Linktypologie etabliert.[243] Eine solche wäre aber durchaus wünschenswert, da es sinnvoll ist, Klassen von Verbindungsstrukturen im Datenbankmodell darzustellen, um eine effektivere Bearbeitung oder Recherche zu ermöglichen.

6.3. Publikationsprozess

6.3.1. Lemmastrecken

An großen Wörterbuchprojekten arbeiten mehrere Generationen von Lexikografen über Jahrzehnte hinweg. Relativ frühzeitig wird ein Wörterbuchplan erstellt, der eine Lemmaliste und ein Artikelstrukturprogramm festlegt. Dieser Plan ist während des Abarbeitens der

[240] Vgl. KRISHNAMURTHY 1987, S. 66ff.
[241] Vgl. KRISHNAMURTHY 1987, S. 67.
[242] Vgl. SCHLAEFER 2002, S. 92ff.
[243] Vgl. HAß-ZUMKEHR 2001, S. 111.

Buchstabenstrecken von A-Z kaum noch zu modifizieren. Zwischen der Bearbeitung der Lexikoneinträge für Lemmata wie *aufmachen* und *zumachen* können mehrere Jahre liegen. Obwohl diese Lemmata morphologische, semantische und etymologische Verwandtschaft zeigen, werden sie von unterschiedlichen Lexikografen in unterschiedlicher Qualität bearbeitet. Es ist daher nicht verwunderlich, dass es in Wörterbüchern mehr Verweise auf frühzeitig erstellte Einträge gibt als umgekehrt.[244] Auch die Wörterbuchbasis verändert sich im Laufe der Bearbeitung, wenn z.B. neue Quellen und Belege die Wissensbestände erweitern.[245] Schließlich erweitert der Lexikograf selbst im Laufe der Arbeit sein Wissen über den Beschreibungsgegenstand und erhöht seine lexikografische Beschreibungskompetenz.[246] Aus diesen Beobachtungen resultiert die Forderung, dass sich der lexikografische Arbeitsprozess an den zu bearbeitenden linguistischen Phänomenen orientieren sollte und nicht an der arbiträren Ordnung des Alphabets.[247]

Digitale Wörterbuchprojekte erlauben in besonderem Maße, den Ablauf der lexikografischen Arbeit nach linguistischen Kriterien auszurichten, da ihre Zugriffsmöglichkeiten gestatten, dass sich die Arbeit an Lemmazeichentypen orientiert, deren syntaktische oder semantische Eigenschaften sich beliebig fein untergliedern lassen.[248] Beispielsweise ist es möglich, dass sich mehrere Forschergruppen an der Eingabe der Datenbanken zu unterschiedlichen lexikologischen Gebieten, die auch Projekt-Module genannt werden können, beteiligen.[249] So lassen sich Projekt-Module z.B. von einer Forschergruppe Synonymik, einer Forschergruppe Fremdwortschatz oder einer Forschergruppe Neologie zur selben Zeit an den gleichen Lemmastrecken bearbeiten.[250]

6.3.2. Front-End-Lösungen

In der Wörterbuchproduktion arbeiten Lexikografen heute mit Dienstprogrammen, die komplexe Prozesse auf einer breiten Datenbasis ermöglichen. PETELENZ beschreibt die Leistungen des von XEROX patentierten Nachschlagesystems LOCOLEX in Fähigkeiten der morphologischen Analyse, Lemmatisierung, Kompositasegmentierung, Erkennung von Mehrwortlexemen und getrennten Präfixverben.[251]

---

[244] Vgl. STORRER 2001, S. 62.
[245] Vgl. LEMBERG 2001, S. 81.
[246] Vgl. LEMBERG 2001, S. 81.
[247] Vgl. STORRER 2001, S. 62.
[248] Vgl. STORRER 2001, S. 63.
[249] Vgl. HAß-ZUMKEHR 2001, S. 103.
[250] Vgl. HAß-ZUMKEHR 2001, S. 107.
[251] Vgl. PETELENZ 2001, S. 210.

Auch die Benutzerschnittstellen zum Verfassen und Editieren der Lexikonartikel sind komplexe Front-End-Lösungen. So hat die Dudenredaktion in den 90er Jahren elektronische Lexikografenarbeitsplätze entwickelt, die Computer unterstützte Lexikografie ermöglichen. Hierbei handelt es sich um ein Programmsystem, das auf verschiedenen Datenbankstrukturen basiert und dem Lexikografen einen SGML-Editor bietet, der nur solche Eingaben zulässt, die nach der Dokumenttypdefinition zulässig sind, womit sichergestellt ist, dass Wörterbucheinträge nur innerhalb der vorgegebenen Mikrostruktur verfasst werden können.[252]
Hinsichtlich der Werkzeuge, mit denen Lexikografen heute an Korpora und Editionen arbeiten, sollte man die Chancen bedenken, die sich ergeben, wenn mehre Forschungsinstitute über das Internet verbunden werden.[253] Hier wird eine räumliche Nähe zum lexikografischen Korpus unerheblich und große Wörterbuchprojekte können prinzipiell ‚peer-to-peer' (im Verbund) bearbeitet werden. Das bedeutet im Idealfall, dass jeder Lemmazeichentyp von jenem Wissenschaftler beschrieben wird, der sich mit entsprechenden Forschungen bereits befasst hat.[254]

6.3.3. TEI-Richtlinien

Um einen Standard für die Auszeichnung von Texttypen aus dem geisteswissenschaftlichen Bereich zu entwickeln, ist 1988 die TEXT ENCODING INITIATIVE (TEI) gegründet worden. Die TEI hat sich 1990 für SGML als Standard-Auszeichnungssprache entschieden.[255] Im Jahr 1994 publiziert sie erstmalig die ‚Guidelines for Electronic Text Encoding and Interchange'[256], die in die aktuell als ‚TEI P4' bezeichneten Richtlinien gemündet sind, wobei ‚P' für Proposal steht und die ‚4' eine Versionsbezeichnung darstellt.[257] In den Richtlinien werden Standards für Dokumentenarchitektur und Dokumenttypdefinitionen ausgeführt. Für Lexika gibt es eigene TEI-Richtlinien, die beanspruchen, Wörterbücher aller „westlicher" Sprachen zu standardisieren.[258] Die TEI-Richtlinien definieren für die grobe Struktur eines gedruckten Wörterbuchs das folgende Tagset (Auszug):

<text> Diese Markierung umschließt jede Art von Texten wie z.B. ein Gedicht oder ein Drama, aber auch Sammlungen von Essays oder einen Roman, bzw. ein Wörterbuch;

<front> umschließt alle einleitenden Texte (Vorworte, Titelseiten, Widmungen);

<body> definiert die zentrale Texteinheit zwischen <front> und <back>;

---

[252] Vgl. WERMKE 1998, S. 53.
[253] Vgl. STORRER 2001, S. 66.
[254] Vgl. STORRER 2001, S. 67.
[255] Vgl. SCHMIDT/MÜLLER 2001, S. 38.
[256] Vgl. SCHMIDT/MÜLLER 2001, S. 38.
[257] Vgl. Website der TEI http://www.tei-c.org
[258] Vgl. SCHMIDT/MÜLLER 2001, S. 38.

<back> definiert Nachworte, Literaturlisten und Anhänge;
** ** beschreibt Untereinheiten der Auszeichnungen <front>, <body>, oder <back>;
<div0> beschreibt die höchste Hierarchieebene einer Untereinheit;
<div1> beschreibt die nächste Untereinheit nach dem <div0> Merkmal;
<entry> umklammert einen regelmäßig strukturierten Lexikonartikel;
<entryFree> bezeichnet einen Lexikonartikel, der keine zwingenden Strukturen beinhaltet;
<superEntry> bezeichnet einen Lexikonartikel, der auf den eines Homografen folgt.
Für alle Tags gelten nur Attribute einer spezifizierten Menge.[259]

Die TEI-Richtlinien sind international anerkannt und werden beispielsweise im DWDS verwendet.[260]

6.4. Blick auf Endprodukte

Im Zusammenspiel von Linguistik und Informatik in der Lexikografie treten zunehmend Validitätskriterien in den Vordergrund, die sich einerseits der qualitativen Beurteilung bereits vorhandener Wörterbücher zuwenden und andererseits einen Kriterienkatalog für neu zu entwickelnde Objekte der Lexikografie anstreben.[261] Die folgenden Ausführungen diskutieren solche Formen der lexikalischen Datenpräsentation im Vergleich zu traditionellen Wörterbüchern, die konzeptionell mit besonders großen Chancen ausgestattet sind.

6.4.1. Online-Lexika im Hypertext

Zwischen Online-Lexika im Hypertextformat und konventionellen Wörterbüchern besteht eine augenfällige Gemeinsamkeit in den Mikrostrukturen, die beiden Medien durch die Organisation von Verweisen ermöglicht, Angaben, Definitionen und Kommentare nicht bloß linear, sondern auch hierarchisch bzw. dreidimensional abzubilden. Allerdings besteht auch ein elementarer Unterschied zwischen beiden Präsentationsmedien gerade in der Ausgestaltung von Mikrostrukturen, da ein online verfügbares Hypertextwörterbuch prinzipiell eine unendliche Menge von Informationsmöglichkeiten antizipieren muss, die in unendlich komplexen Anfragen gesucht werden könnten.[262] STORRER identifiziert weitere wesentliche Unterschiede zwischen den beiden Medien, die sich in etwa wie folgt aufzählen lassen:

[259] Vgl. Website der TEI http://www.tei-c.org
[260] Vgl. Website des DWDS http://www.dwds.de
[261] Vgl. SCHNEIDER 1999, S. 70.
[262] HAß-ZUMKEHR 2001, S. 105 (Hervorhebung im Original).

- Im Gegensatz zu stabilen Wörterbuchartikeln lassen sich die Module des Hypertextes situativ und benutzerspezifisch komponieren.
- Ebenso sind Verweise im Hypertext nicht statisch, sondern durch Programmanweisung dynamisch generierbar.
- Der Zugriff auf Informationen im Hypertext ist ungleich variabler, da die Parameter der Suchabfrage individuell eingestellt werden können.
- Der Hypertext kann nicht nur auf Objekte verweisen, er kann sie durch multimediale Repräsentation regelrecht einbinden.[263]

Aus dieser Aufzählung ergibt sich der wichtigste lexikografische Mehrwert eines online verfügbaren Wörterbuchs gegenüber dem Printmedium, wenn man bedenkt, dass durch das Hypertextkonzept eine Verbindungsstruktur zwischen Datenpräsentation und Datenquelle geschaffen werden kann.[264]

Beispielsweise ist für ein Dialektwörterbuch Multimedialität besonders interessant, da sie eine semasiologische Ausrichtung gestattet, indem Gegenstände bildlich dargestellt werden, deren Bezeichnungen heute nicht mehr bekannt sind, die aber früher zum alltäglichen Wortschatz gehört haben.[265] Ferner kann hier die ohnehin schwierige phonetische Transkription durch Audiostreams von Sprecherproben ergänzt werden.[266]

Das Hypertextlexikon ist somit ganz dicht an den Quellen und Belegen, die das Printprodukt höchstens sporadisch oder in Beispielen angibt. Die große Chance eines Online-Lexikons im Hypertextformat besteht demnach in der Möglichkeit, mit dem Korpus in ständiger Verbindung zu bleiben.

6.4.2. Wörterbuchverbund

Hypertextkonzepte gestatten auch die Integration mehrerer Wörterbuchprojekte in einen Online-Verbund. PETELENZ beschreibt, dass in den 90er Jahren zahlreiche Lexika herausgegeben worden sind, die sich miteinander verknüpfen lassen. Dabei handelt es sich meist um bilinguale Wörterbücher für „*Deutsch-Englisch/Englisch-Deutsch sowie deutsche Fremd- und Synonymwörterbücher - auf CD-ROM in Verbindung mit anderen Nachschlagewerken wie Lexikon, Enzyklopädie, Landkartenatlas. (...) Untereinander verbundene lexikographische Kompendien wie DUDENs LexiRom, Bertelsmanns INFOROM*

[263] Vgl. STORRER 2001, S. 55.
[264] Vgl. LEMBERG 2001, S. 76.
[265] Vgl. LEMBERG 2001, S. 80.
[266] Vgl. LEMBERG 2001, S. 81.

oder Infopedia von Tewi gestatten dem Benutzer den Zugriff auf Informationen aller Werke gleichzeitig, da die Daten zentral in einer Datenbank abgelegt wurden.“[267]

Während solche Wörterbücher für den Massenmarkt bestimmt sind, gibt es inzwischen auch Projekte, die für die Sprachwissenschaft Verbundsysteme lexikalischer Ressourcen anstreben. Hier sei auf den Ansatz hingewiesen, ein digitales deutsches Dialektwörterbuch zu etablieren, das die Einzelprojekte der Territorialwörterbücher in einer modularen lexikografischen Datenbank zusammenführt.[268]

Auch in der diachronen Sprachwissenschaft werden Verbundsysteme mit besonderem Interesse betrachtet. BURCH/FOURNIER beschreiben, wie die Datenbasis zu mhd. Wörtern historisch über mehrere Generationen hinweg schichtweise erarbeitet worden ist und stellen fest: „*BMZ, Lexers Handwörterbuch, seine Nachträge zum Handwörterbuch und das eng auf Lexer bezogene FINDEBUCH, diese vier Wörterbücher müssen also als regelrechter Wörterbuchverbund angesehen werden, dessen stark ausgeprägte Verweisstruktur sich in geradezu idealtypischer Weise für die Abbildung in einer Hypertextstruktur eignet.*“[269] In Göttingen und Trier ist ein elektronisches Text- und Belegarchiv[270] erarbeitet worden, das die zitierten Titel im Projekt ‚Mittelhochdeutsche Wörterbücher auf CD-ROM und im Internet’ bereits in Verbindung setzt.

Hier zeigt sich, dass die Chancen eines Wörterbuchverbunds nicht bloß in der Akkumulation möglichst großer Wissensbestände zu suchen sind, sondern dass sich Lexika wie die Teile eines Puzzles zusammenfügen lassen, aus deren Zusammenspiel sich erst die hinlängliche Beschreibung eines Lexems ergibt.

6.4.3. Lexikalische Datenbanken

Lexikalische Datenbanken sind Wörterbücher, die nicht nur mit Computern hergestellt, sondern auch für Computer bestimmt sind. Sie erfüllen die Anforderungen, welche durch Prozesse maschineller Sprachverarbeitung an lexikalische Ressourcen gestellt werden.[271] Das traditionelle Wörterbuch hat nur eingeschränkte Möglichkeiten, die Zusammenhänge im Wortschatz einer Sprache darzustellen. Es ist im Wesentlichen eine Liste von „Einzelwortschicksalen“.[272] Eine komplex strukturierte Datenbasis, die Beziehungen

[267] PETELENZ 2001, S. 201.
[268] Vgl. FOURNIER 2003.
[269] BURCH/FOURNIER 2001, S. 133ff.
[270] Vgl. PLATE/RECKER 2001, S. 156.
[271] Vgl. MARTIN 1995, S. 4.
[272] GLONING/WELTER 2001, S. 118.

zwischen Wörtern nachempfindet, kann sehr viel mehr sein, wenn sie die Eigenschaften von Wörtern dynamisch in Relation setzt, um semantische Konzepte zu modellieren.[273]

Für lexikalische Ressourcen gibt es inzwischen viele Anwendungsbereiche, die bei KUNZE/WAGNER aufgelistet werden: „*die Maschinelle Übersetzung; die Informationserschließung; die semantische Annotierung von Korpora; die Entwicklung von Sprachlernwerkzeugen, Übersetzungswerkzeugen und Werkzeugen zum Informationserwerb; die Entwicklung automatischer Summarizer; die Realisierung von Sprachgenerierungswerkzeugen.*“[274] Diese Anwendungsbereiche brauchen lexikalische Datenbanken, die neben der bloßen statistischen Beschreibung von Wörtern auch Cluster ihrer Eigenschaften generieren, indem sie Informationen über den gebräuchlichen Kontext von Wörtern und die übliche Verteilung in Texten analysieren.[275]

Nach SCHNEIDER tritt die Computer gestützte Lexikografie aktuell in eine Phase ein, in der das Lexikon zunehmend als Wissensbasis für sprachverarbeitende Systeme betrachtet wird.[276] Von der Qualität lexikalischer Datenbanken wird es zukünftig abhängen, wie gut die jeweiligen Anwendungsbereiche der Sprachtechnologie ihre Probleme lösen.

[273] Vgl. PETELENZ 2001, S. 204.
[274] KUNZE/WAGNER, S. 229ff.
[275] Vgl. BOGURAEV/PUSTEJOVSKY 1996, S. 9.
[276] Vgl. SCHNEIDER 1999, S. 70.

7. Die Benutzerseite

Wenn man der Frage nachgeht, wie aus dem elektronischen Korpus ein Wörterbuch entsteht, sollte die Benutzerseite nicht ignoriert werden, wenngleich hier eher Probleme berührt sind, welche die Lexikografie insgesamt betreffen. Korpusbasierte Lexikografie prägt sich auf der Benutzerseite nicht dadurch aus, dass Wörterbücher nur noch in elektronischer Form verwendet werden. Durch die Unmittelbarkeit, mit der das sprachliche Wissen aus Korpora dem Benutzer auf einem Computerbildschirm zugänglich gemacht werden kann, sollten aber Chancen und Schwierigkeiten diskutiert werden, die durch eine solche Präsentationsform entstehen.

Nutzer sind oft schwer zu überzeugen, das Wörterbuch aus der Hand zu legen, da sie es unbequem finden, den Computer hochfahren zu müssen, um ein Wort nachzuschlagen oder mit dem Buch auch einfach nur zufrieden sind.[277] Es ist daher grundsätzlich nicht sinnvoll, das Printprodukt durch neue Medien ersetzen zu wollen. Es ist davon auszugehen, dass elektronische Wörterbücher eher ergänzenden Zwecken dienen, wenn sie beispielsweise in Textverarbeitungsprogramme integriert sind oder am Arbeitsplatz über das Internet aufgerufen werden können. Hier bieten sie dem Nutzer im Vergleich zum Buch einen gewissen Mehrwert, der im folgenden thematisiert werden soll.

7.1. Textverdichtung

Ein Mehrwert der digitalen Lexikografie gegenüber Printprodukten besteht zunächst für die Lexikografen darin, dass sie nicht mehr durch zu knappen Druckraum gefesselt sind, der Techniken zur Abkürzung und Textverdichtung nötig macht, die sinnvoll sind, um möglichst viele Informationen auf möglichst wenig Papier zu organisieren.[278] Viele Benutzer haben durch diese Techniken der Textverdichtung erhebliche Schwierigkeiten, an Informationen zu gelangen, da sie mit den Abkürzungen, der Nischentechnik und der Tildierung nicht zurecht kommen.[279] Davon abgesehen haben sie aber auch Probleme, die aufgelösten Begriffe in den Benutzungshinweisen oder explizite Kommentare zur grammatischen, semantischen oder pragmatischen Verwendung in den einzelnen Artikeln zu verstehen, weswegen es nötig erscheint, dass sie zukünftig kontextsensitiv in einer Hilfe nachschlagen können.[280]

[277] Vgl. LEMBERG 2001, S. 71.
[278] Vgl. STORRER 2001, S. 56.
[279] Vgl. PETELENZ 2001, S. 211ff.
[280] Vgl. PETELENZ 2001, S. 219.

Hier wird deutlich, dass die Modellierung einer Benutzerschnittstelle für elektronische Wörterbücher nicht mehr im Zusammenhang mit der Modellierung einer Mikrostruktur verstanden werden kann, sondern als eigenständiger lexikografischer Prozess zu sehen ist.[281]

7.2. Verweisstrukturen

Die Verweisstrukturen eines Online-Wörterbuchs bieten gegenüber dem Printprodukt einen großen Mehrwert. Wenn Verweisnester auftreten, ist es für Benutzer wesentlich angenehmer, die Verweishandlung durch anklicken von Hyperlinks statt durch das Bewandern von Bibliotheken vorzunehmen. Durch ein Onlinewörterbuch entfallen demnach zeitliche und räumliche Beschränkungen für die Benutzer.[282] Hyperlinkkonzepte können so auch eine wörterbuchdidaktische Funktion ausüben, da die Benutzer nicht nur der Mühe enthoben sind, internen und externen Verweisen durch aufwendige Prozeduren zu folgen, sondern sich geradezu gereizt fühlen müssen, dem Verweis nachzugehen.[283] Interessant finden es beispielsweise PLATE/RECKER, dass man im Falle des Mittelhochdeutschen durch Hyperlinks die Schwelle zur Benutzung des großen Belegwörterbuchs absenken könnte, da leider immer wieder festgestellt werden müsste, dass „*nur wenige der potentiellen Benutzer den Umstieg vom TASCHENLEXER auf die großen Wörterbücher bewältigen.*“[284]

Einen weiteren Mehrwert, den die Verweisstrukturen eines digitalen Wörterbuchs gegenüber Printprodukten bieten, liegt in der Möglichkeit, multimediale Inhalte in die Artikel einzubinden. STORRER gibt jedoch zu bedenken, dass es nicht reiche, möglichst viele solcher Inhalte mit einem Wörterbuchprojekt zu verknüpfen, um dem Benutzer ein Maximum an Information zu bieten. Wichtig sei vielmehr, dass die Integration unterschiedlicher Zeichen- und Medientypen nach semantisch-funktionalen Prinzipien verlaufe, deren Erforschung im Bereich der Lexikografie aber noch längst nicht abgeschlossen sei.[285]

Die Quantität von Verweisen sollte so bemessen sein, dass ein Maximum an Information ermöglicht, der Nutzer aber nicht überfordert wird.[286] Bei Online-Publikationen ist es grundsätzlich wünschenswert, dass das Nutzerverhalten protokolliert werden kann, um die Verweisstruktur zu verbessern. Interessant wäre es auch, wenn der Nutzer Hyperlinks nach seinen Interessen generieren könnte, was eine Kategorisierung von Verknüpfungsstrukturen dringend erforderlich macht.

---

[281] Vgl. HAß-ZUMKEHR 2001, S. 104.
[282] Vgl. LEMBERG 2001, S. 83.
[283] Vgl. LEMBERG 2001, S. 75.
[284] PLATE/RECKER 2001, S. 168.
[285] Vgl. STORRER 2001, S. 66.
[286] Vgl. HAß-ZUMKEHR 2001, S. 111.

7.3. Zugriffsstrukturen

Während der Nutzer von Printprodukten bisher lernen musste, verdichtete Texte zu lesen, so muss sich der Benutzer digitaler Ressourcen heute ebenfalls die Funktionen der Such- und Filterwerkzeuge aneignen, die ihm einen erweiterten Zugriff auf die gewünschte lexikalische Information gestatten.[287] Prinzipiell ist es so, dass das gedruckte Wörterbuch nur den Zugriff über die alphabetische Ordnung der Lemmata erlaubt,[288] während das elektronische Wörterbuch den freien Zugriff über die Volltextsuche gestattet.[289] LEMBERG spezifiziert weitere Zugriffsmöglichkeiten des digitalen Wörterbuchs: „*Bei den Suchfunktionen (...) handelt es sich im wesentlichen um Volltextsuche, Suchen mit Worttrunkierungen, qualifizierende Suchen mit Hilfe von Wildcards oder Booleschen Operatoren oder durch schreibtolerante oder inkrementelle Suchfunktionen.*“[290] Elektronische Wörterbücher gestatten demnach eine komplexe Recherche anzulegen, die vom Nutzer allerdings auch beherrscht werden muss. Hier ist es jedoch möglich, Wörterbucheinträge auch direkt mit Hilfstexten zu verbinden.[291]

Nutzer von digitalen Wörterbüchern sind darauf angewiesen, dass sie zu einem frühen Zeitpunkt ihrer Suche nicht nur über vorhandene, sondern auch über nicht vorhandene Inhalte informiert werden, was beispielsweise bei Hypertextwörterbüchern durch dynamische Sitemaps geschehen kann, welche die unmittelbare Informationsumgebung detaillierter anzeigen als die weiter entfernt liegende.[292]

STORRER fordert, dass die Benutzerschnittstelle von digitalen Wörterbüchern an Typen von Benutzungssituationen adaptierbar sein müsste.[293] STORRER begründet diese Forderung am Beispiel bilingualer Wörterbücher, wo besonders deutlich werde, dass hier unterschiedliche Muttersprachler in verschiedenen Sprachrichtungen übersetzen, mal nach Synonymen suchen, mal nach Kollokationen und mal nach grammatischen Eigenschaften.[294] PETELENZ bewältigt diese Schwierigkeiten, wenn er Maskierungsmechanismen für sein zweisprachiges Online-Wörterbuch verwendet, um zwischen polnischen und deutschen Benutzern zu unterscheiden, die entweder in die Muttersprache oder in die Fremdsprache übersetzen wollen. Mit Masken

[287] Vgl. STORRER 2001, S. 57.
[288] Vgl. STORRER 2001, S. 53ff.
[289] Vgl. PLATE/RECKER 2001, S. 168ff.
[290] Vgl. LEMBERG 2001, S. 73.
[291] Vgl. LEMBERG 2001, S. 78.
[292] Vgl. HAß-ZUMKEHR 2001, S. 105.
[293] Vgl. STORRER 2001, S. 64.
[294] Vgl. STORRER 2001, S. 64.

sei es möglich, schreibt PETELENZ, durch Selektionen auf der grafischen Oberfläche „*Varianten eines Dokumentes als Untermenge des Gesamtdokumentes zu definieren.*“[295]
Schwierigkeiten treten bei der Konzeption digitaler Wörterbücher bezüglich der Benutzerführung auf, wenn bei der Beschriftung von Schaltflächen, Filtern und Selektoren entschieden werden muss, an welchen Adressatenkreis sich das Wörterbuch richtet.[296] Hier liegt aber auch eine Möglichkeit, künftig einen echten Mehrwert gegenüber dem Buch zu generieren, indem man den Benutzer selbst Profile definieren lässt, die ihn durch das Lexikon führen. Das digitale Wörterbuch liefert hier Vorteile, die von Printprodukten nicht geleistet werden können, da es nicht nur die lexikografischen Beschreibungen, sondern auch die Zugriffsstrukturen an den jeweiligen Kenntnisstand und Informationsbedarf anpassen kann.

7.4. Interaktivität

Der Mehrwert von digitalen Wörterbüchern zeigt sich nicht nur in verbesserten Verweisstrukturen, in verbesserten Zugriffs- und Suchmöglichkeiten, sondern vor allem auch in der Aufhebung der statischen zugunsten einer dynamischen Struktur, wenn Kooperation und Interaktion zwischen Lexikografen und Benutzern ermöglicht wird.[297] Die Lexikografen von Printwörterbüchern haben meist keinen Kontakt zu ihren Lesern und der Benutzerkreis, für den sie das Wörterbuch verfassen, ist ihnen oft unbekannt.[298]
Wörterbücher, die über das World Wide Web für Benutzer und Lexikografen zugänglich sind, erscheinen als besonders interessant, da hier diverse Kommunikationsdienste (E-Mail, Newsgroups, Chat) vorhanden sind, die eine Vielzahl an Möglichkeiten zur Partizipation auf der Nutzerseite bieten. Die aktive Beteiligung der Rezipienten am lexikografischen Prozess reicht in existierenden Online-Wörterbüchern bereits von Hinweisen auf Fehler und Lemmalücken bis hin zu Wörterbüchern, deren Aufbau prinzipiell durch die Beteiligung der Benutzer organisiert ist.[299]
LEMNITZER berichtet von einem Versuch, an dem er beteiligt gewesen ist, um die Nutzung zweisprachiger Online-Wörterbücher zu protokollieren und auszuwerten.[300] In den Zugriffsprotokollen hat er eine erstaunlich hohe Zahl erfolgloser Zugriffe feststellen müssen, die im wesentlichen darin begründet lag, dass ein Suchwort falsch eingegeben worden ist, ein Suchwort im Wörterbuch gefehlt hat oder die Auswahl der Grundform bzw. des Lemmas dem

[295] PETELENZ 2001, S. 210.
[296] Vgl. HAß-ZUMKEHR 2001, S. 109ff.
[297] Vgl. LEMBERG 2001, S. 73.
[298] Vgl. LEMBERG 2001, S. 84.
[299] Vgl. STORRER 2001, S. 66.
[300] Vgl. LEMNITZER 2001, S. 248ff.

Nutzer Probleme bereitet hatte.[301] LEMNITZER führt an gleicher Stelle aus, dass es innerhalb kürzester Zeit möglich gewesen sei, durch einfache Maßnahmen die Zahl erfolgloser Zugriffsversuche drastisch zu senken.

Online-Wörterbücher sind demnach für Lexikografen und Benutzer besonders interessant, weil sie die Aktionen der Rezipienten verfolgen können, wodurch sich Rückschlüsse auf die Qualität des Lexikons ziehen lassen.[302] Wenn es durch Analyse der Zugriffsprotokolle gelingt zu erfahren, welche Benutzungssituationen problematisch sind, können Erkenntnisse dieser Datenerhebungen in Neuauflagen einfließen.[303] Interaktivität ist folglich eine wichtige Voraussetzung für einen Mehrwert gegenüber Printprodukten.

[301] Vgl. LEMNITZER 2001, S. 247.
[302] Vgl. STORRER 2001, S. 67.
[303] Vgl. LEMNITZER 2001, S. 247.

8. Ergebnisse

Im Spannungsfeld zwischen empirischen Verfahren der Sprachdatenerhebung und der Sprachkompetenz des Menschen liegen die größten Chancen und Schwierigkeiten korpusbasierter Lexikografie. Sie potenzieren sich dort, wo die geballte Rechenleistung des Computers der Wortbildungsproduktivität von Sprache gegenübersteht. Korpusbasierte Lexikografie nähert sich dem Wort, dem Sprachzeichen über die Grapheme. Die Beschreibungskompetenz des Lexikografen kann sich Lexemen über die Inhaltsseite der Sprache nähern. Traditionelle und korpuslinguistische Verfahren können sich durch ihre Fähigkeiten nur ergänzen, aber nicht ersetzen.

8.1. Schwierigkeiten

Es ist festzuhalten, dass die Linguistik im Bereich des Korpusdesigns noch mit kleineren Problemen bei der Zusammenstellung der Texte beschäftigt ist, welche vor allem die Textauswahl und die Kategorisierung von Textsorten betreffen. Schwieriger sind die Probleme bei Verfahren der automatisierten Korpusaufbereitung. Die Segmentierung ist ein mühseliger Prozess. Auch bei der Annotierung laufen zum Teil noch aufwendige Prozeduren ab, wenn Korpora morphologisch oder syntaktisch aufbereitet werden. Semantisches Tagging findet fast gar nicht statt und ist vollautomatisch kaum vorstellbar.

Im Bereich der grundlegenden Wissensextraktion gestalten sich Verfahren zur automatischen Lemmatisierung etwas schwerfällig. Verfahren zur spezifischen Wissensextraktion scheitern wiederum meist da, wo Aussagen über die Grundbedeutung von Wörtern getroffen werden sollen. Auch die Extraktion lexikalischen Wissens aus maschinenlesbaren Wörterbüchern zeigt nur kleine Erfolge. Prozesse und Ressourcen zur automatisierten Bedeutungsextraktion konnten bis auf den Ansatz von GERMANET nicht entdeckt werden. Hierin liegt ein ganz großes Defizit der deutschsprachigen Korpuslinguistik. Die Kompilierung von Wörterbüchern wirft dagegen kaum Probleme auf, wenn man davon absieht, dass die umfassende Ausgestaltung der Mikrostrukturen lexikalischer Datenbanken nicht trivial ist. Schwierigkeiten in der Benutzung digitaler Wörterbücher werden, wenn überhaupt, durch die Vielzahl von Möglichkeiten entstehen, die sich bei der Recherche eines Wortes darbieten.

Zusammenfassend kann man sagen, dass sich die meisten Schwierigkeiten auf einem Lösungsweg befinden. Allerdings wird das Kernproblem, die Wortsemantik, bisher nur in einem einzigen Ansatz verfolgt, der nicht erkennen lässt, wohin er sich künftig entwickeln wird.

8.2. Chancen

Die deutsche Korpuslinguistik ist im Vergleich zum angelsächsischen Sprachraum spät gestartet, aber früh erwachsen geworden. Hier haben sich internationale Standards bei der Auszeichnung und Bestimmung von Korpustexten durchgesetzt und interessante Projekte etabliert wie z.B. das WORTSCHATZLEXIKON, das TIGER-Korpus und die COSMAS-Schnittstelle. Das Großprojekt DWDS wird die Chancen korpusbasierter Lexikografie sicherlich noch einmal steigern, sobald es ausgereift ist.

Im Bereich grundlegender Verfahren zur lexikalischen Wissensextraktion reichen die Möglichkeiten zur Gewinnung von Konkordanzlisten und Kollokationsanalysen weit über das hinaus, was man als Standard erwarten könnte. Spezifische Prozeduren zur lexikalischen Wissenserhebung verzeichnen die größten Erfolge bei der Identifikation von Neologismen. Aber auch im Bereich Fachsprache und in der Analyse domänenspezifischer Korpora gibt es gute Ergebnisse. Verfahren zur Erschließung von Semantik werden wiederum die größten Chancen zugerechnet, um sprachtechnologisch gesteuerte Anwendungen voranzutreiben. Hier könnten sich durch die Weiterentwicklung lexikalischer Wissensbasen, z.B. durch Angaben zu Kollokationen und Kohäsionsstärken, große Erfolge feiern lassen, wenn es gelingt, Mehrdeutigkeit zu disambiguieren.

Auch im Bereich der Kompilierung sind hier Chancen vielfach diskutiert worden. Verknüpfungsstrukturen jeglicher Art zwischen Menschen, Objekten und Wissen versprechen den größten Mehrwert gegenüber traditioneller Lexikografie.

Die Chancen korpusbasierter Lexikografie lassen sich darin zusammenfassen, dass sich im deutschen Sprachraum eine gute Infrastruktur herausgebildet hat, um die Eigenschaften von Wörtern zu analysieren. Es kann Jahre dauern, bis alle Möglichkeiten ausgeschöpft sind, welche diese Ressourcen bieten. Ob die beschriebenen Chancen genutzt werden, hängt nicht mehr länger von der Informatik, sondern von der Linguistik ab.

8.3. Ausblick

Wenn man sich die Zusammensetzung deutschsprachiger Korpora betrachtet, stellt man fest, dass ein Wettlauf begonnen hat, der auf die schiere Zahl von Textwörtern zielt. Größe lässt sich aber nur auf Kosten der Repräsentativität erreichen. Korpora wachsen vor allem durch Zeitungen, Fachzeitschriften und Magazine. Es darf bezweifelt werden, dass aus diesen Textsorten beispielsweise eine relevante Aussage zum Wortschatz regionaler Varietäten gewonnen werden könnte. Es besteht die Gefahr, dass Zeitungsredakteure zukünftig die Kategorisierung von Textsorten übernehmen und es ist zu befürchten, dass Korpuslinguisten

es nicht schaffen, komplexe Diskurse wie *Krieg und Frieden* in der Korpusstruktur abzubilden. Die größten Schwierigkeiten und Herausforderungen, die Korpuslinguisten zu bewältigen haben, liegen in der Disambiguierung von Semantik. Probleme bei automatischen Verfahren zur Annotierung, Lemmatisierung, Wissensextraktion und Kompilierung erscheinen vergleichsweise marginal. Wortsemantik kann nur disambiguiert werden, wenn die Inhaltsseite der Sprachzeichen im Korpus repräsentiert ist. Ein Korpus, das nicht nach Jahrgängen, Titeln, Textsorten, sondern nach Diskursen strukturiert ist, wäre für die korpusbasierte Lexikografie von größter Bedeutung.

Verzeichnisse

Abkürzungen

CES Corpus Encoding Standard
DWDS Digitales Wörterbuch der deutschen Sprache
DTD Document Type Definition
EDV Elektronische Datenverarbeitung
IDS Institut für Deutsche Sprache
IKP Institut für Kommunikationsforschung und Phonetik
KWIC KeyWord In Context
LDC Linguistic Data Consortium
LFG Lexical Functional Grammar
MRD Machine Readable Dictionary
OCR Optical Character Recognition
POS Part-Of-Speech
SGML Standard Generalized Markup Language
TEI Text Encoding Initiative
TTR Type-Token-Ratio
XML eXtended Markup Language

Abbildungen

Tabellen

Internetadressen

ACQUILEX (europäisches Projekt zur Lexikonakquisition)
http://www.cl.cam.ac.uk/Research/nl/acquilex/acqhome.html
ANNOTATE (Parser)
http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html
CELEX (lexikalische Ressource)
http://www.kun.nl/celex/
CISLEX (elektronisches Wörterbuch)
http://www.cis.uni-muenchen.de/projects/cislex.html
DEPARTMENT OF COMPUTATIONAL LINGUISTICS AND PHONETICS (Universität des Saarlandes)
http://www.coli.uni-sb.de
DEUTSCHES RECHTSWÖRTERBUCH (historisches Online-Wörterbuch)
http://www.rzuser.uni-heidelberg.de/~cd2/drw/index.htm
DEUTSCHES WÖRTERBUCH von Jacob und Wilhelm Grimm (online)
http://www.dwb.uni-trier.de/index.html
DWDS (Korpus und elektronisches Wörterbuch)
http://www.dwds.de
EAGLES (Expertengruppe für Auszeichnungsstandards für Korpora)
http://www.ilc.cnr.it/eagles/home.html
ELWIS (Projekt zur korpusbasierten Entwicklung lexikalischer Ressourcen)

http://www.sfs.nphil.uni-tuebingen.de/elwis/elwisinfo.html

EUROWORDNET (Europäisches Projekt zu semantischen Netzen)

http://www.sfs.nphil.uni-tuebingen.de/lsd/

GENIOS (kommerzielles Archiv)

http://www.genios.de

GESELLSCHAFT FÜR LINGUISTISCHE DATENVERARBEITUNG

http://www.gldv.org

GERMANET (lexikalisch-semantisches Netz)

http://www.sfs.nphil.uni-tuebingen.de/lsd/

INSTITUT FÜR DEUTSCHE SPRACHE, Mannheim

http://corpora.ids-mannheim.de

INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG (Universität Stuttgart)
http://www.ims.uni-stuttgart.de

KLASSIKERWORTSCHATZ (korpusbasiertes historisches Wörterbuch)

http://www.klassikerwortschatz.uni-freiburg.de

LIMAS (Korpus)

http://linux-s.ikp.uni-bonn.de/Limas/index.htm

LINGUISTIC DATA CONSORTIUM

http://www.ldc.upenn.edu

MITTELHOCHDEUTSCHE WÖRTERBÜCHER AUF CD-ROM UND IM INTERNET

http://www.mwv.uni-trier.de/index.html

NEGRA (Korpus)

http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html

PARGRAM (Grammatikprojekt)

http://www.ims.uni-stuttgart.de/projekte/pargram/ (deutsche Sektion)

http://www2.parc.com/istl/groups/nltt/pargram/ (internationale Sektion)

SIMPLE (Internationales Projekt zur semantischen Annotierung)

http://www.ub.es/gilcub/simple/simple2.html

TEXT ENCODING INITIATIVE (Standardisierung von Textauszeichnung)

http://www.tei-c.org

TIGER (Korpus)

http://www.ims.uni-stuttgart.de/projekte/tiger/

UNICODE CONSORTIUM (Standardisierung von Zeichensätzen)

http://www.unicode.org

WORLD WIDE WEB CONSORTIUM (Standardisierung von Auszeichnungssprachen)

http://www.w3.org

WORTSCHATZLEXIKON (Korpus und elektronisches Wörterbuch)
http://wortschatz.uni-leipzig.de

Literatur

ARGENTON, HANS: Indexierung und Retrieval von Feature-Bäumen am Beispiel der linguistischen Analyse von Textkorpora. Sankt Augustin 1998

BIBER, DOUGLAS/ CONRAD, SUSAN/ REPPEN, RANDI: Corpus linguistics - Investigating language structure and use. Cambridge 1998

BÖHME, TIMO/ RAHM, ERHARD: XML-Datenbanksysteme. Architekturen und Benchmarks. In: Bullinger, Hans-Jörg/ Weisbecker, Anette (Hrsg.): Content Management - Digitale Inhalte als Bausteine einer vernetzten Welt. Stuttgart 2002, S. 1-14

BOGURAEV, BRANIMIR/ PUSTEJOVSKY, JAMES: Issues in Text-based Lexicon Acquisition. In: Boguraev, Branimir/ Pustejovsky, James (Hrsg.): Corpus Processing for Lexical Acquisitions. Cambridge/ London 1996, S. 3-17

BREIVIK, LEIV EGIL/ HASSELGREN, ANGELA (Hrsg.): From the COLT's mouth ... and others': language corpora studies in honour of Anna-Brita Stenström. Amsterdam/ New York 2002

BÜCHEL, GREGOR/ SCHRÖDER, BERNHARD: Verfahren und Techniken in der computergestützten Lexikografie. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg.): Chancen und Perspektiven computergestützter Lexikografie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001 S. 7-28

BULLINGER, HANS-JÖRG/ WEISBECKER, ANETTE (Hrsg.): Content Management - Digitale Inhalte als Bausteine einer vernetzten Welt. Stuttgart 2002

CLEAR, JEREMY: Computing. In: Sinclair, John M. (Hrsg.): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London 1987, S. 41-61

ELSEN, HARALD/ HARTMANN, JÖRG: Satzbandanalyse und Aufbereitung des Definitionswortschatzes eines deutschen Wörterbuchs. In: Haller, Johann/ Pütz, Horst P. (Hrsg.): Sprachtechnologie: Methoden, Werkzeuge, Perspektiven. Vorträge im Rahmen der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Kiel, 3.-5. März 1993. Hildesheim 1993, S. 169-181

ENGELBERG, STEFAN/ LEMNITZER, LOTHAR: Lexikographie und Wörterbuchbenutzung. Tübingen 2001

FELLBAUM, CHRISTIANE: Wordnet. Cambridge 1998

FOURNIER, JOHANNES: Vorüberlegungen zum Aufbau eines Verbundes von Dialektwörterbüchern. In: Zeitschrift für Dialektologie und Linguistik 70. Stuttgart 2003, S. 155-176.

GLONING, THOMAS/ WELTER, RÜDIGER: Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe-Wörterbuch. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 117-133

GRÜN, ANGELA VON DER: Wort-, Morphem- und Allomorphhäufigkeit in domänenspezifischen Korpora des Deutschen. Erlangen 1999

HAß-ZUMKEHR, ULRIKE: Zur Mikrostruktur im Hypertextwörterbuch. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 103-116

HEYER, G./ QUASTHOFF, U./ WOLFF, C.: Möglichkeiten und Verfahren zur automatischen Gewinnung von Fachbegriffen aus Texten. In: Bullinger, Hans-Jörg/ Weisbecker, Anette (Hrsg.): Content Management - Digitale Inhalte als Bausteine einer vernetzten Welt Stuttgart 2002, S. 43-49

HEYN, MATTHIAS: Zur Wiederverwendung maschinenlesbarer Wörterbücher (Lexicographica Series Maior 45). Tübingen 1992

KAMMER, MANFRED: Korpora geschriebener Sprache. In: Lenders, Winfried (Hrsg.): Computereinsatz in der angewandten Linguistik : Konstruktion und Weiterverarbeitung sprachlicher Korpora. Frankfurt am Main 1993, S. 49-62

KATZ, JERROLD J./ FODOR, JERRY A.: Die Struktur einer semantischen Theorie. In: Steger, Hugo: Vorschläge für eine strukturale Grammatik des Deutschen. Darmstadt 1970, S. 202-268

KINNE, MICHAEL: Der lange Weg zum deutschen Neologismenwörterbuch. In: Teubert, Wolfgang (Hrsg.): Neologie und Korpus. Tübingen 1998, S. 63-110

KRISHNAMURTHY, RAMESH: The Process of Compilation. In: Sinclair, John M. (Hrsg.): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London 1987, S. 62-85

KÜNNETH, THOMAS: Datenbankgestützte Speicherung von Korpora. Erlangen 2001

KUNZE, CLAUDIA/ WAGNER, ANDREAS: Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In: Lemberg, Ingrid/ Storrer, Angelika / Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 229-246

LANGER, STEFAN: Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX. CIS-Bericht-96-94. München 1995

LEMBERG, INGRID: Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 71-92

LENDERS, WINFRIED: Semantische Relationen in Wörterbuch-Einträgen - Eine Computeranalyse des Duden-Universalwörterbuchs. In: Schaeder, Burkhard (Hrsg.): Lexikon und Lexikographie. Vorträge im Rahmen der Jahrestagung 1990 der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Siegen, 26.-28. März 1990. Hildesheim u.a. 1990, S. 92-105

LENDERS, WINFRIED (Hrsg.): Computereinsatz in der angewandten Linguistik: Konstruktion und Weiterverarbeitung sprachlicher Korpora. Frankfurt am Main 1993(a)

LENDERS, WINFRIED: Tagging - Formen und Tools. In: Haller, Johann/ Pütz, Horst P. (Hrsg.): Sprachtechnologie: Methoden, Werkzeuge, Perspektiven. Vorträge im Rahmen der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Kiel, 3.-5. März 1993. Hildesheim 1993(b), S. 369-392

LENDERS, WINFRIED/ WILLÉE, GERD: Linguistische Datenverarbeitung. Opladen 1998

LENZ, SUSANNE: Korpuslinguistik (Studienbibliographien Sprachwissenschaft Bd. 32). Tübingen 2000

LEZIUS, WOLFGANG: Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Stuttgart 2002

LJUNG, MAGNUS: What vocabulary tells us about genre differences: A study of lexis in five newspaper genres. In: Breivik, Leiv Egil, Hasselgren, Angela (Hrsg.): From the COLT's mouth ... and others' : language corpora studies in honour of Anna-Brita Stenström. Amsterdam/ New York 2002, S. 181-196

MANNING, CHRISTOPHER D./ SCHÜTZE, HINRICH: Foundations of statistical natural language processing (2. überarbeitete Aufl.). Cambridge/ London 2000

MARTIN, WILLY: Maschinelle Lexikografie: Ein Blick in die Zukunft. In: Hitzenberger, Ludwig (Hrsg.): Angewandte Computerlinguistik, Hildesheim u.a. 1995, S. 1-21

MÜLLER, CAROLIN/ SCHMIDT, INGRID: Entwicklung eines lexikographischen Modells: Ein neuer Ansatz. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 29-52

NEUMANN, ROBERT: Korpora – Eine Herausforderung an die Informationserschließung. In: Feldweg, Helmut (Hrsg.): Lexikon und Text. Tübingen 1996, S. 13-36

PETELENZ, KRYSZTOF: Das Informationsdesign auf der Speicherungsebene eines zweisprachigen Online-Wörterbuchs Polnisch-Deutsch. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 199-228

PLATE, RALF/ RECKER, UTE: Elektronische Materialgrundlage und computergestützte Ausarbeitung eines historischen Belegwörterbuchs. Erfahrungen und Perspektiven am Beispiel des neuen Mittelhochdeutschen Wörterbuchs. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 155-179

SCHAEDER, BURKHARD: Lexikon und Lexikographie. Vorträge im Rahmen der Jahrestagung 1990 der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Siegen, 26.-28. März 1990. Hildesheim u.a. 1990

SCHLAEFER, MICHAEL: Lexikologie und Lexikographie. Eine Einführung am Beispiel deutscher Wörterbücher. Berlin 2002

SCHNEIDER, RENÉ: Maschineller Erwerb lexikalischen Wissens aus kleinen und verrauschten Textkorpora. München 1999

SINCLAIR, JOHN M.: Corpus Concordance Collocation. Oxford 1991

SINCLAIR, JOHN M.: Korpustypologie. Ein Klassifikationsrahmen. In: Teubert, Wolfgang (Hrsg.): Neologie und Korpus. Tübingen 1998, S. 111-128

STORRER, ANGELIKA: Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/ Storrer, Angelika/ Schröder, Bernhard (Hrsg): Chancen und Perspektiven computergestützter

Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen 2001, S. 53-70

TELLENBACH, ELKE: Neologismen der neunziger Jahre. Vom Textkorpus zur Datenbank. In: Barz, Irmhild/ Fix, Ulla/ Lerchner, Gotthard (Hrsg.): Das Wort in Text und Wörterbuch. Leipzig 2001

TEUBERT, WOLFGANG: Korpus und Neologie. In: Teubert, Wolfgang (Hrsg.): Neologie und Korpus. Tübingen 1998, S. 129-170

WAUSCHKUHN, OLIVER: Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. Aachen 1999

WEBER, NICO: Computergestützte Analyse von Definitionstexten in einem deutschen Wörterbuch. In: Haller, Johann/ Pütz, Horst P. (Hrsg.): Sprachtechnologie: Methoden, Werkzeuge, Perspektiven. Vorträge im Rahmen der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Kiel, 3.-5. März 1993. Hildesheim 1993, S. 140-168

WERMKE, MATTHIAS: Vorüberlegungen zum Aufbau elektronischer Textkorpora in der Dudenredaktion. In: Bergmann, Rolf (Hrsg.): Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung, 1. und 2. November 1996. Stuttgart/ Leipzig 1998, S. 49-56

WILLÉE, GERD: Erfahrungen mit morphologischem Tagging am Beispiel des LIMAS-Korpus. In: Haller, Johann/ Pütz, Horst P. (Hrsg.): Sprachtechnologie: Methoden, Werkzeuge, Perspektiven. Vorträge im Rahmen der Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V., Kiel, 3.-5. März 1993. Hildesheim 1993, S. 352-368

ZIERL, MARCO: Entwicklung und Implementierung eines Datenbanksystems zur Speicherung und Verarbeitung von Textkorpora. Erlangen 1998

Erklärung

Hierdurch erkläre ich, dass ich meine Hausarbeit zur Erlangung des Magister-Grades (M.A.):

Korpusbasierte Erstellung eines Wörterbuchs des Deutschen

Chancen und Schwierigkeiten

selbständig ohne unerlaubte Hilfe verfasst, ganz oder in Teilen noch nicht als Prüfungsleistung vorgelegt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die anderen Quellen im Wortlaut oder dem Sinn nach entnommen wurden, sind durch Angabe der Herkunft kenntlich gemacht.

Marburg, den 08. September 2003

_____

Marc Meyer