Automatische Wortschatzerschließung großer Textkorpora am Beispiel des DWDS

Alexander Geyken (Berlin)

Abstract

In the past years a large number of electronic text corpora for German have been created due to the increased availability of electronic resources. Appropriate filtering of lexical material in these corpora is a particular challenge for computational lexicography since machine readable lexicons alone are insufficient for systematic classification. In this paper we show – on the basis of the corpora of the DWDS – how lexical knowledge can be classified in a more fine-grained way with morphological and shallow syntactic parsing methods. One result of this analysis is that the number of different lemmas contained in the corpora exceeds the number of different headwords of current large monolingual German dictionaries by several times.

1 Einleitung

Bei der Frage nach der Zugehörigkeit eines Wortes zum Wortschatz der deutschen Sprache beschränken sich Wörterbücher, auch die Großwörterbücher, bewusst. Um in ein Wörterbuch aufgenommen zu werden, muss ein Wort über längere Zeit von mehreren Sprechern, am besten in mehreren Textsorten, mit einer gewissen Häufigkeit verwendet werden. Es sollte sich auch nicht völlig einfach aus den Wortbestandteilen erschließen lassen, also nicht semantisch transparent sein, zumindest jedoch "usualisiert" sein oder eine orthographische Besonderheit (z. B. *Ich-AG*) aufweisen. In diesem Zusammenhang stellt sich jedoch die Frage, auf welcher Beleggrundlage diese Häufigkeiten ermittelt werden können, und ob es nicht wichtige Wörter geben kann, die in ihrer Häufigkeit unter dem gewählten Schwellwert liegen oder nicht in allen Textsorten vorkommen. Darüber hinaus ist zu klären, wie die reine Quantität der Beleggrundlage mit der Menge an darin vorkommenden verschiedenen Wörtern zusammenhängt. Ist es so, dass ein Textarchiv ab einer gewissen Größe alle Wörter des deutschen Wortschatzes umfasst, zumindest all diejenigen, die heutzutage im Gebrauch sind, oder findet man bei der Betrachtung von immer größeren Textmengen auch immer mehr neue Wörter?

Noch nie in der Geschichte der Wortschatzforschung war es möglich, diesen Fragen in derselben empirischen Breite nachzugehen. In den letzten 15 Jahren wuchs die Anzahl an digital verfügbaren Texten nahezu exponentiell. Einen besonders großen Anteil daran haben Zeitungen, die wie kein anderes Druckerzeugnis digital erhältlich sind, sowie im World Wide Web verbreitete Texte, die nur noch digital verbreitet werden. Gleichzeitig verbessern sich die technologischen und computerlinguistischen Voraussetzungen, sehr große Textmengen maschinell durchsuchbar zu machen.

Stützten sich die Wörterbuchmacher noch bis vor wenigen Jahren auf manuell erstellte Exzerpte von Wortbelegen in einem Umfang von ein paar tausend, allenfalls, wie im Falle des größten deutschen Wörterbuchs, dem Grimmschen Wörterbuch, ein paar Millionen Belege, so enthält heute jedes große Zeitungsarchiv mehrere Millionen Dokumente und somit mehrere

Milliarden potentieller Wortbelege. Hiermit eine systematische Wortschatzarbeit zu betreiben, ist eine Herausforderung nicht nur für die Wörterbuchredaktionen der Verlage, sondern auch für die Wissenschaft. Es geht dabei, wie in der Folge gezeigt wird, um sehr große Wortbestände, die einerseits in der Praxis nicht "per Kopf" durchgesehen werden können, andererseits aber auch der maschinellen Auswertung Probleme bereiten. Zur Illustration sei lediglich ein Beispiel aufgeführt. Ein Ziel der Wortschatzarbeit ist die zuverlässige Entdeckung von Neologismen, wie beispielsweise das durch das soziale Netzwerk *Twitter* in Gebrauch gekommene Verb *twittern* oder das im Zuge der Arbeitsmarktreform eingeführte Substantiv *Aufstocker*. Diese Wortformen sind vor ihrer "Entdeckung" gerade eben nicht in elektronischen Lexika enthalten. Maschinelle Verfahren könnten Neologismen lediglich als unbekannt markieren und würden diese somit fälschlicherweise mit Eigennamen, fremdsprachlichen Wörtern oder einfach Rechtschreibfehlern auf eine Stufe setzen. Darüber hinaus haben maschinelle Verfahren das Problem, dass sie semantisch transparente und damit vergleichsweise "langweilige" Wortzusammensetzungen nicht von den "interessanten" unterscheiden können. Beispielsweise würden Wörter wie *Zimmertür*, *Wohnungstür*, *Badezimmertür* oder *Esszimmertür* mit entsprechenden Wortbildungsregeln genauso in ihre Wortbestandteile zerlegt wie das in den letzten Jahren durch eine EU-Richtlinie wieder in die Schlagzeilen gekommene Wort *Selbstabfertigung*, obwohl dessen spezifische Bedeutung, also der Einsatz von Personal von Reedern oder Fluggesellschaften beim Be- und Entladen der Schiffe bzw. Flugzeuge, sich nicht aus der Bedeutung der Einzelwörter herleiten lässt. Da die Zahl der zerlegbaren Wortformen im Deutschen sehr groß ist, sinkt mit dieser Gleichbehandlung von transparenten und nicht-transparenten Formen die Wahrscheinlichkeit, dass ein interessantes Wort inmitten der uninteressanten Bildungen von den Lexikografen gefunden wird. In der Tat ist im konkreten Fall die *Selbstabfertigung* nicht in den beiden aktuellen deutschen Großwörterbüchern, Duden und Wahrig[1], enthalten.

In diesem Beitrag möchte ich auf der Grundlage des an der Berlin-Brandenburgischen Akademie erstellten Textarchivs des Projekts DWDS einige der computerlinguistischen Verfahren vorstellen, die gegenwärtig eingesetzt werden, um den in den Texten enthaltenen Wortschatz systematisch auszuwerten. Diese Verfahren stellen Vorarbeiten für die in der zweiten Phase des Projekts beginnende lexikografische Arbeit dar, indem sie den Lexikografen bei der Sichtung der großen Materialfülle unterstützen sollen.

2 DWDS-Kernkorpus und das erweiterte Korpus

Das Textkorpus des DWDS wurde, unterstützt von der Deutschen Forschungsgemeinschaft, in den Jahren 2000–2003 erstellt und wird seitdem kontinuierlich ausgebaut. Es setzt sich aus zwei großen Bestandteilen zusammen: dem kleineren nach Textsorten ausgewogenen, öffentlich recherchierbaren Kernkorpus sowie dem im Wesentlichen auf neueren Zeitungsquellen fußenden Ergänzungskorpus.

Das Kernkorpus soll den Wortschatz des gesamten 20. Jahrhunderts in größtmöglicher Ausgewogenheit widerspiegeln. Es wurde daher darauf geachtet, die Textsorten über das gesamte Jahrhundert gleichmäßig zu streuen und die prozentuale Verteilung der Textsorten untereinander angemessen zu berücksichtigen. Das Kernkorpus umfasst etwa 100 Millionen Textwörter; dies entspricht in etwa einer kleinen Bibliothek von ca. 1'500 Monographien. Etwa 40% davon wurden in mehr als zweieinhalb Jahren Arbeit mit bis zu 20 studentischen Mitarbeitern digitalisiert, der Rest wurde von Verlagen angekauft bzw. von Textgebern eingeworben. Aufgenommen wurden Dokumente aus fünf Bereichen: Schöne Literatur: 27%; Journalistische Prosa: 26%; Fachprosa 22%; Gebrauchstexte: 20%; transkribierte Texte gesprochener Sprache: 5%. Bei der Auswahl wurde das Projekt von Mitgliedern der Berlin-

[1] Duden Universalwörterbuch, 6. Auflage, 2006; WAHRIG.digital Deutsches Wörterbuch, 2003.

Brandenburgischen Akademie der Wissenschaften beraten; eine gewisse Zufälligkeit bei der Auswahl herrscht lediglich im Bereich "gesprochene Sprache", wo Daten vor der zweiten Jahrhunderthälfte kaum verfügbar sind. Die Bereiche setzen sich nach folgenden Auswahlkriterien zusammen:

Schöne Literatur: darunter wird nicht nur die "hohe Literatur" verstanden, sondern auch die Unterhaltungsliteratur, die bislang lexikografisch kaum aufgearbeitet worden ist. Unter dem Aspekt eines breiten Nutzerkreises sind Konsalik und Höber nicht minder wichtig als Grass und Strittmatter. Pro Dekade enthält das Korpus etwa 20 längere Prosawerke (v. a. gehobene Literatur, aber auch Erzählungen für Kinder- und Jugendliche, literarische Tagebücher, etc.) sowie 10 Werke der Unterhaltungsliteratur, wobei der Übergang vom "Groschenroman" zum Unterhaltungsbestseller und zur gehobenen Literatur naturgemäß fließend ist.

Journalistische Prosa: Diese Textsorte umfasst sowohl die überregionalen Tages- und Wochenzeitungen, aber auch einige regionale Blätter, die unter lexikografischen Aspekten oft besonders interessant sind; weiterhin Magazine und Illustrierte, unter Einschluss der "gelben Presse" und von Jugendzeitschriften. Zeitungen bilden keine homogene Textsorte; das Feuilleton ist anders als der Wirtschaftsbericht, die Sportseite anders als die Kleinanzeigen. Die Auswahl erfolgte sowohl ereignisorientiert als auch seriell. Die aufwendig anmutende Auswahl der Zeitungsausgaben nach historischen Ereignissen beruht auf der Erfahrung, dass bestimmte Ausdrucksweisen im Zusammenhang mit solchen Ereignissen geläufig geworden sind, z. B. für 1900 der 12.11. (Ende der Pariser Weltausstellung), 1901 der 10.12. (erste Verleihung des Nobelpreises), 1902 der 31.5. (Ende des Burenkriegs). Im seriellen Zugriff wurde für die jeweilige Zeitung eine gewisse Anzahl von Ausgaben für jedes Jahr zufällig ausgewählt. Im Einzelnen umfasst das Korpus u. a. eine Auswahl der Berliner Zeitungen (Vossische Zeitung, Berliner Tageblatt), zusätzlich wahlweise bzw. nach Verfügbarkeit eine Nummer aus der Frankfurter, Kölner und Münchner Tagespresse. Aufgenommen wurde darüber hinaus für jedes Jahr jeweils eine Nummer einer Wochenzeitung bzw. Magazins: für die Nachkriegszeit Die ZEIT, Der Spiegel; für die Zeit davor Berliner Illustrierte bzw. Neue Berliner Illustrierte, Die Gartenlaube oder der Simplizissimus.

Fachprosa: Hier wurden aus einer Reihe von Fachgebieten, von Philosophie und Jurisprudenz, über Medizin und Theologie bis zu Chemie, Physik und Mathematik, maßgebliche Texten dieses Jahrhunderts aufgenommen. Diese umfassen sowohl Aufsätze aus wissenschaftlichen Zeitschriften wie auch wissenschaftliche Monographien; angestrebt wurde hier ein ungefähres Gleichgewicht zwischen den verschiedenen Disziplinen.

Gebrauchstexte: dies ist eine Gruppe von Texten, die in der Wörterbucharbeit nur selten berücksichtigt werden – Gebrauchsanweisungen, Beipackzettel, Theaterprogramme, Werbetexte. Aufgenommen wurden pro Dekade je ein Kochbuch, ein Gesundheitsratgeber, ein Reiseführer, ein Benimm- oder Familienhausbuch, eine technische Dokumentation, 10 Gebrauchsanleitungen bzw. Beipackzettel, Werbetexte (aus den berücksichtigten Zeitungs- und Magazinausgaben), ferner sämtliche juristische Texte aus den in der Jurisprudenz allgemein verwendeten Sammlungen "Schönfelder" und "Sartorius".

Transkriptionen gesprochener Sprache: dieses Teilkorpus umfasst Transkriptionen sowohl spontan gesprochener als auch nicht spontan gesprochener Sprache des 20. Jahrhundert. Insbesondere sind darin enthalten: Redensammlungen (u. a. von Kaiser Wilhelm, Hitler, Ulbricht und Honnecker), Rundfunkansprachen von 1929–1944, die in Kooperation mit dem Deutschen Rundfunkarchiv transkribiert wurden, österreichische Parlamentsprotokolle (1948–1956) sowie Bundestagsprotokolle der 14. Sitzungsperiode. Es umfasst Auszüge aus dem Projekt *Emigrantendeutsch in Israel* (Anne Betten, Universität Salzburg) und dem Literarischen Quartett (1988–2001). Schließlich enthält das Korpus Transkripte von Interviews, u. a. aus Spiegel und ZEIT, aber auch des *Berliner Wendekorpus* (Norbert Dittmar, FU-

Berlin), einem Korpus narrativer Interviews von knapp 100 West- und Ostberlinern, die zum Ereignis des 9. November befragt wurden.

Das DWDS-Ergänzungskorpus ist wesentlich größer als das DWDS-Kernkorpus: es umfasst über 1.5 Milliarden Textwörter, genügt jedoch nicht den Kriterien der Ausgewogenheit, sondern besteht im Wesentlichen aus elektronisch verfügbaren Zeitungsquellen der letzten 20 Jahre, die auch in Print-Ausgaben erschienen und somit bibliographisch referenzierbar sind. Vorwiegend wurden überregionale Tageszeitungen und Zeitschriften aufgenommen, darunter *Bild*, die *Frankfurter Allgemeine Zeitung*, *Neue Zürcher Zeitung*, *Süddeutsche Zeitung*, *taz* und *Die Welt*. Ebenfalls sind *Spiegel*, *ZEIT* und *Konkret* enthalten. Darüber hinaus wurden die auf Berlin bezogene *Berliner Zeitung* und der Berliner *Tagesspiegel* aufgenommen.

Grundlage der im Folgenden dargestellten Wortschatzstudie sind das gesamte DWDS-Kernkorpus (100 Millionen Textwörter) sowie der indexierte Teil des DWDS-Ergänzungskorpus (gegenwärtig ca. 900 Millionen Textwörter). Zusammengenommen besitzen beide Textcorpora einen Umfang von über 1 Milliarde aneinandergereihter Textwörter, die im Folgenden der Einfachheit halber als DWDS-Korpus bezeichnet werden.

3 Exkurs: Korpora und Wörterbücher – ein quantitativer Vergleich

Die Größe von Korpora lässt sich sowohl anhand der aneinandergereihten Wörter (Tokens) als auch nach der Zahl der verschiedenen Wörter (Types) messen. In Tabelle 1 wird der Begriff Token in einem naiven Sinne, nämlich als Zeichenkette zwischen zwei Leerstellen charakterisiert. Die Zahlen zeigen zumindest zweierlei. Auf der einen Seite wird sichtbar, dass das Deutsche gegenüber dem Englischen eine größere Zahl von Types bei gleicher Tokenanzahl aufweist. Dies lässt sich vergleichsweise einfach dadurch erklären, dass die Wortbildung und vor allem die Komposition im Deutschen produktiv sind, d. h. potentiell unendlich viele Wörter gebildet werden können. Mit statistischen Methoden lässt sich nachweisen, dass diese produktiven Mechanismen der Sprache auch genutzt werden: d. h. dass die Anzahl der verschiedenen Wörter in den Korpora nicht gegen eine feste Obergrenze geht, sondern mit zunehmender Textmenge wächst (Geyken 2008).

Korpusname	**Anzahl Tokens**	**Anzahl Types**
Brown Corpus	1 Million	50.000
Limas Korpus	1 Million	110.000
British National Corpus (BNC)	100 Millionen	650.000
DWDS-Kernkorpus	100 Millionen	2.2 Millionen
DWDS-Gesamtkorpus	1 Milliarde	9 Millionen

Tabelle 1: Anzahl von Tokens und Types ausgewählter Korpora[2]

Wörterbuch	**Anzahl Stichwörter**
Wörterbuch der Deutschen Gegenwartssprache (WDG)	88'000
Duden	200'000
Grimm	300'000

Tabelle 2: Anzahl der Stichwörter ausgewählter deutscher Großwörterbücher

[2] Zahlen für das BNC, das Brown- und Limas-Korpus, vgl. Hausser 1998;
für das DWDS-Korpus: *www.dwds.de*.

Der Frage nach dem Vergleich von Stichwortanzahl in Wörterbüchern und Korpusgröße soll im Folgenden anhand einiger bekannter Korpora nachgegangen werden. Vergleicht man die Zahlen in Tabelle 1 und Tabelle 2, so fällt unmittelbar auf, dass das nach den Kriterien des Brown-Corpus in den siebziger Jahren erstellte deutsche Pendant, das Limas-Korpus, mit seinen 110'000 Types viel zu klein ist, um mit der Stichwortanzahl von Wörterbüchern konkurrieren zu können. Die Größenverhältnisse kehren sich bei dem hundert Mal größeren DWDS-Kernkorpus um. Denn die 2.2 Millionen verschiedenen Wörter sind weit mehr als die Stichwortanzahlen großer Wörterbücher. Noch einmal um mehr als das Vierfache erhöht sich diese Zahl beim DWDS-Gesamtkorpus. Dieses enthält etwa neun Millionen verschiedene Wortformen. Im Vergleich dazu muten das *Wörterbuch der deutschen Gegenwartssprache* (WDG) mit 88'000 Stichwörtern, der 10-bändige Duden mit etwa 200'000 und selbst das größte deutsche Wörterbuch, die Erstausgabe des *Wörterbuchs von Jacob Grimm und Wilhelm Grimm*, mit seinen 300'000 Stichwörtern, klein an.

Aus der Anzahl der verschiedenen Wortformen alleine lässt sich jedoch nicht einfach ableiten, dass Korpora einen immensen Schatz an lexikografischem Material enthalten, der weit über die in Wörterbüchern beschriebene Stichwortanzahl hinausgeht. Das liegt daran, dass die Stichwortanzahl in Wörterbüchern etwas anderes misst als die Anzahl der verschiedenen Types in Korpora. Insbesondere werden bei der oben erwähnten Zählweise, derzufolge Wörter Zeichenketten zwischen zwei Leerzeichen sind, auch Kardinalzahlen, Typenbezeichnungen oder Eigennamen mitgezählt. Ebenso finden sich fremdsprachliche Wortformen darunter, da Bücher, Zeitschriften und Zeitungen zuweilen fremdsprachige Zitate enthalten. Der weitaus größte Anteil des lexikografisch uninteressanten Materials besteht aber in der großen Menge transparenter Komposita. Beispielsweise enthält das DWDS-Korpus zu *Tür* die transparenten und damit lexikografisch vergleichsweise uninteressanten *Holztür*, *Stahltür*, *Badtür*, *Schlafkammertür*, *Schlafzimmertür*, *Stalltür*, *Stubentür*, *Wohnzimmertür*, *Wohnungstür*, *Zimmertür*, etc[3]. Es stellt sich somit die Frage, ob Korpora nach Abzug all dieser Wortformen noch lexikografisch interessantes Material enthalten. Der Nachweis darüber, dass sehr große Korpora auch Wörterbuchlücken im großen Maßstab aufdecken können, ist darüber zu führen, wie die unbekannten und gleichermaßen lexikografisch interessanten Wörter aus diesen großen Textmengen effektiv extrahiert werden können.

4 Korpora als Grundlage für die Erstellung von Wörterbüchern

Noch Mitte der 1990er Jahre wurde darauf verwiesen, dass Korpora nicht alle Stichwörter eines Wörterbuchs enthalten und somit in Vielfalt hinter den Belegarchiven der Lexikographen zurückbleiben. Beispielsweise schreibt Hausser zum Verhältnis des Webster's mit dem ausgewogenen, 100 Millionen Textwörter großen British National Corpus:

> Darüber hinaus gibt es viele Wörter im Webster's, die im BNC kein einziges Mal belegt sind, z. B. *aspheric*, *bipropellant*, *dynamotor* – trotz seiner Größe und des Bemühens um ein repräsentatives, balanciertes Korpus. Somit kann die Type-Liste eines großen Korpus zwar helfen, ein traditionelles Lexikon zu ergänzen. Es ist jedoch nicht zu erwarten, dass sich ein großes Korpus als ebenso vollständig oder vollständiger als ein traditionelles Lexikon erweist (Hausser 1998: 5).

In den letzten 10 Jahren hat sich die Situation gegenüber der von Hausser (1998) beschriebenen Lage gewandelt. Die derzeit in den großen Wörterbuchverlagen eingesetzten Korpora sind um eine Größenordnung gewachsen und umfassen nun wenigstens eine oder sogar mehrere Milliarden aneinandergereihte Textwörter. Damit enthalten sie nicht nur weitaus mehr verschiedene Wortformen (vgl. Tabelle 1), womit die oben genannten Begrenzungen

[3] Zwar lassen sich diese Wörter aus ihren Einzelbedeutungen mühelos herleiten. Sie werden dennoch in den Großwörterbüchern verzeichnet, da sie aufgrund ihrer Gebrauchshäufigkeit als usualisiert gelten.

bezüglich der Vollständigkeit weitestgehend entfallen. Diese neue Größenordnung führt darüber hinaus dazu, dass statistische Verfahren besser greifen und somit, wenn sie mit computerlinguistischen Methoden gekoppelt sind, sehr effiziente Filtermethoden darstellen, um die aus den Korpora extrahierten Wortlisten mit den Stichwortlisten der Korpora abzugleichen. In der Tat dürfte es kaum mehr ein großes einsprachiges Wörterbuch geben, welches nicht korpusbasiert aktualisiert wird. Korpora gehören zum Standardwerkzeug für die Neubearbeitung aller großen einsprachigen Wörterbücher in nahezu allen Bereichen der Makro- und Mikrostruktur von Wörterbuchartikeln (z. B. Heid 2004, Klosa 2007).

Auf der Makroebene spielen Textkorpora bei der Bewertung von Stichwörtern eine Rolle. Als Kriterium hierfür lassen sich die Frequenz wie auch die Streuung der Belege über die Zeit und über Textsorten hinweg einsetzen: durch den Abgleich der lemmatisierten Formen des Korpus mit dem Stichwortinventar lassen sich diejenigen Stichwörter extrahieren, die gar nicht oder nur selten belegt sind. Ab hier kommt die lexikografische Arbeit ins Spiel: Fehlende oder neue Wörter können ergänzt, veraltete und überflüssige Begriffe können gestrichen werden.

Auf der Ebene der Mikrostruktur werden bereits seit längerem phraseologische Einheiten in Wörterbüchern auf der Grundlage von Textkorpora neu bearbeitet. Darüber hinaus lassen sich morphologische Angaben wie Genus und Numerus auf der Grundlage von Korpora bewerten, bei geeigneten Korpusfiltermethoden lässt sich dies auf die Rektionsangaben ausweiten. Schwieriger gestaltet sich die Neubewertung von Registerangaben, obwohl auch für diesen Bereich korpuslinguistische Arbeiten vorliegen (z. B. Biber 1994).

Vom methodischen Zugang am einfachsten ist die Überprüfung von Stichwörtern in Bezug auf die Vollständigkeit der Derivations- und Kompositionsmuster. Hierzu setzt man zu einem Stichwort gängige Derivations- und Kompositionsregeln ein und gleicht sie mit dem Korpus ab. Beispielsweise überprüft man für den Eintrag *antichambrieren*, ob die möglichen Bildungen *Antichambrist* oder *Antichambrierer*, *Antichambrierung*, *antichambrierbar* etc. im Korpus belegt sind. Analog dazu würde man bei Komposita das Erstglied des Kompositums nehmen und mit allen Wortformen im Korpus abgleichen, die mit dem Erstglied beginnen. Beispielsweise würde man zur Suche nach allen Belegen von *Mörtel* die Wortformen *Mörteleimer*, *Mörtelfuge*, *Mörtelgeruch*, *Mörtelkorn*, *Mörtelkalk*, *Mörtelklecks*, *Mörtelputz*, *Mörtelspur*, *Mörtelwerk* etc. im Korpus finden und mit denen, die im Wörterbuch aufgeführt sind, abgleichen.

Im nächsten Abschnitt soll der Frage nachgegangen werden, wie computerlinguistische Verfahren eingesetzt werden, um die relevanten lexikalischen Formen aus den sehr großen Korpora zu ermitteln und auf welche Grenzen sie stoßen.

5 Automatische Erschließung der Lexik

Wie bereits erwähnt, ist davon auszugehen, dass die im DWDS-Gesamtkorpus enthaltenen 9 Millionen verschiedenen Wortformen nur zu einem Teil aus lexikografisch interessanten Wortformen bestehen. Wie lassen sich diese Formen extrahieren und vor allem klassifizieren? Bei der Extraktion verfährt man in etwa so wie die gängigen Programme zur Rechtschreibprüfung. Lexikografisch relevant sind zunächst alle Wortformen, die in maschinellem Abgleich mit einem elektronischen Lexikon als gültige Wortformen erkannt oder von einem Wortzerlegungsprogramm auf gültige Lexikoneinträge zurückgeführt werden können. Umgekehrt wird mit speziellen Programmen zur Erkennung von Eigennamen und sonstigen Sonderformen (z. B. Datumsangaben oder Typenbezeichnungen etc.) eine Negativliste mit lexikografisch uninteressantem Material erstellt. Die Negativliste hat zum Ziel, die Menge der nicht analysierbaren Wortformen zu minimieren, um somit den Suchraum für Neologismen bzw. Wörtern, die für die maschinelle Analyse bisher unbekannt sind, möglichst klein zu halten. Schließlich wird eine wortübergreifende Filterkomponente eingesetzt, um Mehrwort-

verbindungen wie *a priori* oder *Runder Tisch* zu extrahieren, oder auch zusammengesetzte Verben, die in der Verbzweitstellung aus zwei Wortbestandteilen, dem Verbstamm und einem Verbpräfix, bestehen. Im Folgenden wird die maschinelle Extraktion der Listen näher erläutert.

5.1 Erstellung der "Positivliste"

In die Positivliste werden alle potentiell relevanten lexikografischen Wortformen aufgenommen. Zur Erkennung dieser Wortformen bedienen wir uns eines automatischen Analyseprogramms, welches versucht, beliebige Wortformen mit Hilfe großer Lexika und einem umfangreichen Regelinventar auf sogenannte Stammformen zurückzuführen, damit jedes Wort unabhängig von der flektierten Form nur ein einziges Mal aufgenommen wird. Beispielsweise werden die flektierten Wortformen *Arzt*, *Ärzte*, *Arztes* oder *Ärzten* auf ein und dieselbe Stammform, nämlich *Arzt*, zurückgeführt. Zusammengesetzte Wortformen wie beispielsweise *Taschenbücher* oder *Landesgruppenvorsitzenden* werden in ihre Bestandteile (*Tasche*, *Buch* bzw. *Land*, *Gruppe*, *Vorsitzende*) zerlegt und auf ihre Stammformen im Nominativ Singular, also *Taschenbuch* bzw. *Landesgruppenvorsitzende* abgebildet.

Für die Wortformenzerlegung wird das TAGH-Morphologiesystem (Geyken/Hanneforth, 2006) eingesetzt. Das TAGH-System besteht aus mehreren Stammlexika: 85'000 Nomen, 23'000 Verben, 18'000 Adjektiven, 2'700 Adverbien, sowie 500 Wortformen aus dem Bereich der geschlossenen Formen. Darüber hinaus sind darin erfasst: 136'000 geografische Namen, 242'000 Familiennamen, 6'000 Organisationen. Schließlich enthält es Sonderformen, darunter ein Teillexikon von Abkürzungen (20'000), von Akronymen (12'000) und von Fremdsprachlichen Material (1'000 Einträge). Bei der Wortanalyse kommen mehr als 1'000 Wortbildungsregeln zum Einsatz. Damit können nicht nur Nominalkomposita in ihre Bestandteile zerlegt werden, sondern auch wortartenübergreifende Wortbildungen erkannt werden. Ein Beispiel hierfür ist die Bildung von Adjektiven aus einem verbalen Stamm, indem man das Suffix -bar hinzufügt, wie in *verhandelbar* oder *heilbar*. Andere Wortbildungsregeln bilden aus einem Ortsnamen und der Endung -er Einwohnernamen. Hierdurch wird beispielsweise *Donaueschinger* auf *Donaueschingen* zurückgeführt, indem die -en Endung getilgt und durch eine -er Endung ersetzt wird.

Das TAGH-Lexikon enthält ferner etwa 12'500 Mehrwortlexeme, die dadurch charakterisiert sind, dass sie nicht diskontinuierlich sind, durchaus jedoch Binnenflexion erlauben. Bei knapp 10'000 Lexemen handelt es sich um geographische Eigennamen wie *New York*, *Rocky Mountains* oder *Rote Meer* (mit den weiteren Formen *Rotes Meer*, *Roten Meer(es)*). Weitere knapp 2'000 Einträge sind zusammengesetzte Nomen-Nomen oder Adjektiv-Nomen Verbindungen, die meisten davon Eigennamen, wie z. B. *Fashion Week*, *Olympische Spiele* oder *Goldene Bär*; aber auch Appellativa wie *Goldenes Buch*, *Aloe Vera* oder *Dolce Vita*. Schließlich umfasst das Lexikon komplexe Adverbien wie *in flagranti* und *a priori*. Eine etwas andere Schwierigkeit besteht in der Erkennung von Präfixverben, wie beispielsweise *vorkommen* oder *auftauchen*. In der Verbzweitstellung sind diese Lexeme separiert, d. h. der Verbstamm steht vor dem Präfix, welches die Phrase beendet. Beispielsweise werden in dem Satz *Die Unterlagen tauchten wieder auf*, die Wortformen *tauchen* und *auf* dem Lexem *auftauchen* zugeordnet. Im Projekt DWDS wird dies durch eine flache Grammatikkomponente (Part-of-Speech Tagger) realisiert (Jurish 2003), die aufgrund statistischer Häufigkeiten entscheidet, ob es sich bei einer Wortform, die rechts von einem Verb steht, um eine Präposition handelt oder um ein abtrennbares Präfix, welches dem Verb zuzuordnen ist. Nicht erfasst werden diskontinuierliche oder variable Mehrwortausdrücke, die Gegenstand des ebenfalls an der BBAW beheimateten Projekts *Kollokationen im Wörterbuch* unter der Leitung von Christiane Fellbaum waren (Fellbaum 2006, 2007), im engeren Sinne jedoch nicht der lexikalischen Extraktion zuzurechnen sind.

Das TAGH-Morphologiesystem erreicht bei neueren Zeitungstexten eine Erkennungsrate von über 99%. Die Genauigkeit der Analyse wurde bislang noch nicht systematisch ausgewertet. Einige grundlegende Prinzipien werden jedoch von der Zerlegungskomponente der TAGH-Morphologie unterstützt. Insbesondere wird gewährleistet, dass nicht alle Zusammensetzungen, die potentiell zerlegt werden können, auch zerlegt werden. Beispielsweise darf der *Gendarm* nicht in *Gen* und *Darm* zerlegt werden, genauso wenig wie der *Eisenhut*, bei welchem die Pflanzenbedeutung durch die Zerlegung in *Eisen* und *Hut* verloren geht.

Andere Wortformen lassen sich auf mehrere Arten zerlegen. In diesem Fall besteht das Ziel der maschinellen Analyse darin, die richtige Zerlegung zu identifizieren und darüber hinaus jede Zerlegung auf den korrekten Stamm zurückzuführen. Das Wort *Telekommunikation* beispielsweise wird in der maschinellen Analyse auf vier verschiedene Weisen zerlegt. Das Symbol '#' steht hierbei für die Wortgrenze: Tele#kommunikation, Tele#komm#unikat#ion, Tele#komm#uni#kation und Telekom#muni#kation (Muni = schweiz. Zuchtstier). Hier ist nur die erste Analyse plausibel, was in der maschinellen Analyse des TAGH-Systems – etwas vereinfacht ausgedrückt – dadurch gewährleistet wird, dass die Zerlegung mit den wenigsten Wortgrenzen als die plausibelste Analyse angesehen wird. Dies ist eine in vielen, aber nicht in allen Fällen ausreichende Heuristik. Beispielsweise kann *Wochenarbeitstag* in Wochen#Arbeits#Tag und in Wochen#Arbeit#Stag zerlegt werden. Beide Analysen weisen eine gleiche Anzahl von Zerlegungen auf. Zur Eliminierung der falschen Analyse bedient man sich einer weiteren Heuristik: die Wortform *Stag* (ein in der Seemannssprache verwendeter starker Draht zum Sichern und Stützen von Masten in der Längsrichtung des Schiffes) taucht in Textkorpora nicht nur selten auf, sondern ist dort auch niemals Teil eines Kompositums. Aus diesem Grunde werden alle Kompositaanalysen mit *Stag* ausgeschlossen, wodurch im Beispiel *Wochenarbeitstag* nur die korrekte Analyse übrigbleibt.

Es gibt jedoch auch Fälle, die kontextlos auf Wortebene nicht eindeutig zerlegt werden können und bei der die automatische Stammzerlegung an ihre Grenzen stößt. Beispielsweise lässt sich eine Wortform wie *Ministern* auf zwei Arten zerlegen, einerseits als Pluralform von *Minister*, andererseits als kleiner Himmelskörper, als Mini#Stern. In diesem Beispiel sind beide Zerlegungen morphologisch möglich und sogar semantisch plausibel. In den folgenden Beispielen hingegen sind automatische Verfahren jedoch problematisch. Die Wortform *Tauschwert* kann entweder in die Stämme Tausch#Wert oder als Tau#Schwert zerlegt werden oder *Kegelschnitte* können als Kegel#Schnitt oder als Kegel#Schnitte analysiert werden. In beiden Beispielen ist nur die erste Analyse sinnvoll, die zweite jedoch semantisch unsinnig. Für die lexikografische Arbeit stellt diese Ambiguität kein Problem dar, da beide Wörter in gängigen einsprachigen Wörterbüchern enthalten sind (z. B. Duden-Universalwörterbuch, Wörterbuch der deutschen Gegenwartssprache, Wahrig-Online). Aufgrund der gegenüber der Anzahl der Stichwörter in Wörterbüchern erheblich größeren Anzahl an verschiedenen Types in den Korpora kann im allgemeinen Fall jedoch nicht davon ausgegangen werden, dass ambige Formen in den Wörterbüchern bereits beschrieben sind. Beispiele für Wörter, die nicht im Duden-Universalwörterbuch aufgeführt, aber bei der automatischen Stammformenanalyse Probleme bereiten, sind die Wortformen *Tauschrate* und *Gewürzdosen*, die als Tausch#Rate oder als Tau#Schrat bzw. als Gewürz#Dose bzw. Gewürz#Dosis zerlegt werden können. In beiden Fällen stellt die automatische Analyse keine Hilfe dar, sondern eine lexikografische Analyse "per Kopf" muss nachgeschaltet werden.

Das Ergebnis dieser maschinellen Analyse ist zweierlei: erstens eine Liste von zerlegbaren Wortformen, die auf ihre Stammformen zurückgeführt sind, und zweitens eine Ordnung der Zerlegungen nach ihren jeweiligen Stämmen, wobei einige Wortformen mehrere Zerlegungen aufweisen können. Es ist somit möglich, diesen Bestand lexikografisch zu untersuchen und mit dem Wörterbuchbestand abzugleichen. Beispielsweise könnten alle Wortformen untersucht werden, die die Bestandteile *Tausch* oder *Schrat* enthalten. In ersterem Fall würde man

die lexikografisch interessante Bildung *Tauschrate* finden, in letzterem das falsche Tau#Schrat, welches man nach lexikografischer Durchsicht eliminieren würde. Ebenso wäre es möglich, die Produktivität von Ableitungen zu untersuchen, beispielsweise wie häufig Adjektive mit -bar gebildet werden, oder welche Substantive den Suffix -itis bilden, wie beispielsweise *Telefonitis*. Schließlich dient diese Liste dazu, einen Überblick über das in deutschen Texten tatsächlich verwendete Vokabular zu gewinnen.

5.2 Erstellung der "Negativliste"

Mit speziellen Programmen werden zunächst alle Wortformen herausgefiltert, die aus Zahlenkombinationen, Datumsangaben oder aus Buchstaben-Zahlenkombinationen bestehen. Hierbei handelt es sich beispielsweise um Formen wie 1.2.2006, 7:4 oder 030-2037-0, S65. Darüber hinaus werden spezielle Bindestrich-Kombinationen oder Schrägstrich-Kombinationen gefiltert: *20er-Liga*, *ZX-4*, *3/10*, *3/100stel*, *afp/dpa*, *Björndalen/Norwegen*. Schließlich werden Abkürzungsformen wie beispielsweise, *str*. wie in *Rumfordstr*. oder *pl*. wie in *Bordeauxpl*. identifiziert und in die Negativliste aufgenommen. Gelegentlich werden Wortformen in Texten ausschließlich groß- bzw. kleingeschrieben. Diese werden normalisiert. Sind sie dann dem Lexikon unbekannt, werden sie als Akronyme klassifiziert.

Die wichtigste Komponente bei der Erstellung der Negativliste stellt die Erkennung der Eigennamen dar. Personen-, Orts- oder Unternehmensnamen, ebenso auch Produktnamen gehören nicht zum lexikografisch interessanten Material, und sie lassen sich in den Texten aufgrund ihrer Kontexte im Satz identifizieren. Die Funktionsweise der Eigennamenerkennung soll an folgenden Beispielen illustriert werden.

(1) *Der Werkzeugmacher Martin Bartensteyn freut sich.*
(2) *Bartensteyn hatte schließlich Feierabend.*
(3) *Der Vollblutpolitiker Ami Ajalon.*
(4) *Problemflüchtling Stoiber – Zurück in die weiß-blaue Gemütlichkeit.*
(5) *Erwin Teufel, der ehemalige Ministerpräsident von Baden-Württemberg.*
(6) *Der berühmte Schewi, Ministerkollege Hans-Dietrich Genschers.*
(7) *Klaus Sterns Dokumentarfilm gewann den Preis.*
(8) *Die Rocky Mountains haben den schönsten Pulverschnee Amerikas.*
(9) *Der Stahlriese Defasco weigert sich, eine Stellungnahme abzugeben.*

Im engen Kontext von Eigennamen finden sich häufig Personen- oder Unternehmensbezeichnungen wie z. B. *Werkzeugmacher*, *Ministerpräsident*, *Politiker*, *Stahlriese* oder *Tochterunternehmen*. In appositiver Stellung charakterisieren sie Funktionen oder Relationen zu Personen bzw. Unternehmen. Dies wird von einer im Projekt DWDS eingesetzten Grammatikkomponente genutzt (Didakowski et al. 2007), um unbekannte Wortformen als Personen bzw. Unternehmensnamen zu identifizieren (Beispiele (1), (2), (3), (5) und (9)). Die Grammatikkomponente beruht auf einer großen Liste von Vor- und Nachnamen, geographischen Namen, Unternehmensnamen sowie einer Liste von etwa 25'000 Personenbezeichnern. Durch die morphologische Komponente (s. unten) wird diese Liste noch einmal vergrößert, indem nicht in dieser Liste enthaltene Personenbezeichnungen wie *Vollblutpolitiker* oder *Problemflüchtling* in ihre Bestandteile, d. h. Vollblut#Politiker oder Problem#Flüchtling zerlegt und somit auf in der Liste enthaltene Personenbezeichnungen zurückgeführt werden. Die Personenerkennung ist nicht satzbezogen, sondern erfolgt textweise. Dadurch ist es möglich, unbekannte Wortformen ohne ausreichenden Kontext als Eigennamen zu klassifizieren. Dies geschieht beispielsweise in (2). Hier wird davon ausgegangen, dass der Name *Bartensteyn* in (1) in einem größeren Kontext eingeführt wurde, so dass der Name ab dann dem System bekannt ist, und in einem zweiten Textdurchlauf als

Eigenname klassifiziert werden kann. Ferner wird das System auch für die Auflösung mehrdeutiger Personennamen eingesetzt. Beispielsweise sind die Wortformen *Stern* oder *Teufel* entweder ein Appellativum oder ein Eigenname. In den Kontexten (5) und (7) hingegen weisen die engen Kontexte diese Wortformen als Namen aus. Die Erkennung von Personennamen bei mehrdeutigen Wortformen geschieht im Rahmen dieser Analyse zu dem Zweck, korrekte Häufigkeitslisten zu erzeugen. Aus diesem Grunde wird auch die Erkennung von Mehrwortausdrücken in die Eigennamenerkennung integriert, um zu vermeiden, dass Kombinationen wie *Rocky Mountains* als eine Vornamen-Namen Kombination im Genitiv identifiziert werden.

6 Ergebnis der Filterprozeduren

Mit Hilfe der morphologischen Analyse werden etwa 6 Millionen der 8.9 Millionen Wortformen des DWDS-Gesamtkorpus als potentiell lexikografisch relevant klassifiziert, also der Positivliste zugeordnet. Die 6 Millionen erkannten Wortformen wiederum können auf 3.9 Millionen Stammformen reduziert werden. Zur Illustration der Materialfülle, die daraus entsteht, haben wir die im 10-bändige Großwörterbuch des Dudens (1999) aufgeführten Komposita, die mit *Selbst-* beginnen, mit dem im Korpus dazu extrahierten Wortformen verglichen. Der Duden verzeichnet 244 Einträge von *Selbstabholer* über *Selbstbedienung* und *Selbsterfahrung* bis hin zu *Selbstzweifel*. Demgegenüber enthält das DWDS-Gesamtkorpus 10'934 verschiedene Selbstkomposita (Types), die mit Hilfe der TAGH-Morphologie auf 7'180 verschiedene Lemmata abgebildet wurden. Der Abgleich der hochfrequenten Selbstkomposita mit den Stichwörtern des Duden ergibt, dass alleine 56 Wortformen, die mehr als 100 Mal im Korpus vorkommen, nicht im Duden verzeichnet sind. Bei den Lemmata der Häufigkeit zwischen 30 und 99 gibt es sogar 251 Selbstkomposita, die nicht in den Duden aufgenommen wurden[4]. Nicht alle Komposita der Liste sind lexikalisiert. Ein kurzer Blick auf die Liste offenbart jedoch schnell einige gravierende Lücken: Die *Selbstauskunft* sollte in einer Neubearbeitung des Duden ebenso wenig fehlen wie die *Selbstbedienungsmentalität*, der *Selbstmordanschlag*, das *Selbstmordattentat* und *-attentäter* oder die *Selbstregulierung*.

Der Negativliste können mit Hilfe der Vorverarbeitungsregeln und des Eigennamenfilters etwa 1.2 Millionen verschiedene Wortformen als lexikografisch nicht interessantes Material zugeordnet werden.

Es verbleiben somit knapp 1.7 Millionen verschiedene Wortformen, die bislang nicht zugeordnet werden konnten. Diese Liste enthält, wie eine exemplarische Durchsicht einiger hundert Wortformen zeigt, vornehmlich Eigennamen, regionale Varianten (wie *Ick*, *nit*, *wa* etc.), orthographische Fehler, fremdsprachliches Material (*mon*, *the*, *que*), ungebräuchliche Abkürzungen (*stellvertr.*, *Kammerorch.*) Wörter, die der historischen Rechtschreibung vor 1902 folgen (*diktiren*, *Litteratur*, *Antheil*), sowie in einem geringeren Maße lexikografisch relevantes Wortmaterial, welches für die maschinelle Analyse unbekannt ist und somit noch aufgearbeitet werden muss.

7 Ausblick

In den letzten 15 Jahren haben sich durch das nahezu exponentielle Wachstum von verfügbaren elektronischen Textressourcen völlig neue Perspektiven für die Wortschatzforschung eröffnet, da mit Textkorpora und den darauf angewandten computerlinguistischen Methoden ein weitaus effektiverer Blick auf den Wortschatz ermöglicht wird. Automatische Verfahren zur Identifikation neuer Wortformen können – je größer die Korpora, desto bessere Ergebnisse können statistische Verfahren liefern – die Vorauswahl nicht nur auf eine reprä-

[4] Die vollständige Liste zusammen mit Belegbeispielen findet sich unter *www.dwds.de/pages/pages_textba/selbst.html*.

sentativere Basis stellen, sondern diese auch massiv beschleunigen. Bei der Neubearbeitung von Wörterbüchern gehören korpusbasierte Methoden der Wortschatzbearbeitung mittlerweile zum Standard in nahezu allen Bereichen der Makro- und Mikrostruktur von Wörterbuchartikeln. Die lexikografische Kompetenz können diese Methoden auf absehbare Zeit jedoch nicht ersetzen. Die Entscheidung darüber, ob es sich bei einer neuen Wortform um eine lexikalisierte Form oder lediglich eine transparente Bildung handelt, muss weiterhin von Lexikografen getroffen werden. Die Basis für eine systematische Wortschatzanalyse bleibt somit nach wie vor eine semantische und wird sich auf absehbare Zeit einer vollständig automatischen Analyse entziehen.

Literatur

Biber, Douglas (1994): "Representativeness in corpus design". In: Zampolli, Antonio/ Calzolari, Nicoletta/Palmer, Martha (eds.): *Current Issues in Computational Linguistics: In Honour of Don Walker*. Dordrecht, Kluwer: 377–407. (= *Linguistica Computazionale* IX–X).

Didakowski, Jörg/Geyken, Alexander/Hanneforth, Thomas (2007): "Eigennamenerkennung zwischen morphologischer Analyse und Part-of-Speech Tagging. Ein automatentheoriebasierter Ansatz". *Zeitschrift für Sprachwissenschaft* 26/2: 157–186.

Fellbaum, Christiane (ed.) (2006): *Corpus-Based Studies of German Idioms and Light Verbs*. Special issue of the *Journal of Lexicography* 19/4.

Fellbaum, Christiane (ed.) (2007): *Collocations and Idioms. Corpus-Based Linguistic and Lexicographic Studies*. Birmingham, UK: Continuum.

Geyken, Alexander/Hanneforth, Thomas (2006): "TAGH. A complete morphology for German based on weighted finite state automata". In: Yli-Jyrä, Anssi/Karttunen, Lauri/ Karhumäki, Juhani (eds.): *Proceedings of Finite-State Methods in Natural Language Processing (FSMNLP) 2005, Lecture Notes in Artificial Intelligence (LNAI)* 4002. Berlin/Heidelberg, Springer: 55–66.

Geyken, Alexander (2008): "Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus". In: Cori, Marcel et al. (eds.): *Construction des faits en linguistique: la place des corpus*. Paris, Larousse: 77–94. (= *Langages* 171).

Hausser, Roland (1998): "Häufigkeitsverteilung deutscher Morpheme". *LDV-Forum* 15/1: 6–26.

Heid, Ulrich et al. (2004): "Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition". *LREC 2004 Proceedings*. Lissabon: 911–914.

Jurish, Bryan (2003): *A Hybrid Approach to Part-of-Speech Tagging*. Final report, Project 'Kollokationen im Wörterbuch', Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf (Stand: September 2009).

Klosa, Annette (2007): "Korpusgestützte Lexikographie besser schneller". In: Kallmeyer, Werner/Zifonun, Gisela (eds.): *Sprachkorpora: Datenmengen und Erkenntnisfortschritt*. Walter de Gruyter: Berlin.