Alexander Geyken
Manuskript
Erscheint in Lili: Heft 136

Korpora als Korrektiv für einsprachige Wörterbücher

1. Einleitung

Waren früher Belegarchive in Form von Zettelkästen die einzige maßgebliche Grundlage für die Aktualisierung von Wörterbüchern, so treten heute zunehmend große elektronische Textkorpora an deren Stelle. Auch wenn diese (noch) nicht in der Lage sind, die in den Zettelkästen enthaltenen menschlichen Sprachbeobachtungen vollständig zu ersetzen, so leisten Textkorpora bereits jetzt schon wertvolle Dienste bei der Neubearbeitung großer Wörterbücher. In der Tat wird heute kaum ein großes einsprachiges Wörterbuch in den großen Kultursprachen ohne die Hilfe von großen elektronischen Textkorpora neubearbeitet. Im deutschen Sprachraum gilt dies für den Duden sowie die Neubearbeitung des Wahrig ebenso wie für die Neubearbeitung des „Deutschen Wörterbuchs“ von Jacob Grimm und Wilhelm Grimm. Folgendes Beispiel aus der Süddeutschen Zeitung vom 2. Juli 2002 motiviert die Verwendung großer Korpora:

> „Eigentlich hätte uns das Wort nicht durch die Lappen gehen dürfen“, gesteht Beate Varnhorn, Chefredakteurin von Wahrig, der Wörterbuchmarke bei Bertelsmann. Doch immer wieder ist das Wort *Ceranfeld* Sprachbeobachtern durchgerutscht. Diese durchforsten im Auftrag der Wörterbuchmacher gedruckte Texte nach so genannten Neologismen. Nun bekommen die Fahnder Unterstützung von Computerlinguisten. Deren Programme sollen Texte schneller nach neuen Wörtern durchsuchen und dabei weniger Fehler machen. Vor zwei Jahren begann man bei Wahrig mit der Entwicklung eines solchen Systems“
>
> Süddeutsche Zeitung, Dienstag, 2.7.2002, Seite V2/7, Wissenschaft

Die Idee, die hinter diesem Vorgehen steckt, beruht auf der Annahme, dass die Sammlung sehr großer Textmengen in digitaler Form mit vergleichsweise wenig Aufwand möglich ist sowie darauf, dass man mit Suchmaschinen die Texte vollständig, d.h. für jedes in dem Text enthaltene Wort durchsuchen kann. Bevor ich auf die neuen Anwendungsmöglichkeiten dieser neuen Form der automatischen Belegexzerption eingehe, soll diese zunächst mit der klassischen Form der Belegsammlungen, den Zettelkästen, verglichen werden.

2. Automatische Belegexzerption und Zettelkästen

Belegsammlungen im traditionellen Sinne sind zunächst einmal Sammlungen von Textausschnitten, die in der Regel von mehreren Personen aus Büchern oder Zeitschriften exzerpiert wurden. Bei diesem Vorgehen werden – zumindest bei den großen Vorhaben – eine größere Anzahl von Personen beauftragt, interessante Wörter oder Verwendungsweisen von Wörtern aus den verschiedensten Quellen – im Falle der Erstausgabe des Deutschen Wörterbuchs von Jacob Grimm und Wilhelm Grimm, etwa 25.0000 Quellen[1] – zu sammeln. Diese werden dann als Karteikarte in einem Zettelkasten oder aber in neuerer Zeit als

[1] Diese 25.000 Quellen sind die Grundlage des Belegarchivs; sie wurden jedoch nicht systematisch exzerpiert.

elektronische Karteikarte in einer Datenbank abgelegt. Durch die Konzentration auf besondere Auffälligkeiten wird gewährleistet, dass die Anzahl der Belege pro Stichwort in einem überschaubaren Rahmen bleibt. Trotz dieser die Beleganzahl reduzierenden Vorauswahl können Belegarchive auf mehrere Millionen Belege anwachsen. Beispielsweise umfasst die Belegsammlung der Neubearbeitung des Deutschen Wörterbuchs alleine für die Buchstaben A-C etwa 3 Millionen Belege, die, verteilt auf mehr als 1000 Schubladen, 20 Regalmeter einnehmen. Die manuelle Belegexzerption hat gegenüber der vollautomatischen Belegexzerption zwei Nachteile: die mangelnde Vollständigkeit – Wörter können übersehen werden – und die Konsistenz der Exzerption, da nicht jede Person gleich exzerpiert und je nach Suchauftrag sogar ein- und dieselbe Person zu verschiedenen Zeitpunkten unterschiedlich exzerpiert.

Auch ein elektronisches Textkorpus, welches die Texte komplett aufnimmt, lässt sich als Archiv von Belegsammlungen begreifen. Nur trifft in diesem Falle die Suchmaschine, die den Exerpierer ersetzt, keine Vorauswahl, sondern indiziert alle Wörter und Wortverbindungen. Somit entspricht jedes Wort des Texts dem Stichwort der Karteikarte und das Textwort zusammen mit dem Kontext dem Beleg der Karteikarte. Diese vollständige Exzerption hat im Unterschied zur manuellen Textexzerption den Vorteil, dass kein Wort des „gelesenen“ Texts „übersehen“ werden kann. Auf der anderen Seite stellt sich natürlich die Frage, inwieweit es möglich ist, Texte in elektronischer Form in einer ähnlichen qualitativen Streuung zu erwerben, wie dies in den großen traditionellen Belegarchiven der Fall ist. Dieser Frage nach der Größe und der Streuung wird im nächsten Abschnitt nachgegangen. Darüber hinaus wirft die vollständige Belegexzerption das Problem auf, dass durch Erfassung jedes einzelnen Wortes gewaltige Belegmengen entstehen, die ohne Hilfsmittel nahezu unmöglich durchgesehen werden können. Ein Beispiel soll verdeutlichen, in welche Dimensionen man hier vorstoßen kann. Nehmen wir dazu an, dass ein relativ kleiner Roman, wie das Treibhaus von Wolfgang Koeppen, automatisch exzerpiert werden soll. Dieser Roman umfasst 51.810 aneinandergereihte Wortformen, darunter befinden sich 11.879 graphemisch verschiedene Wortformen. Jede dieser Wortformen stellt ein potentielles Stichwort dar. Einigen Wortformen des Romans, wie z.B. *und* (1780 Vorkommen) oder *Keetenheuve* (680) werden sehr viele Belege zugeordnet, andere Wortformen, z.B. *Treibhaus* oder *Antichambrierer* tauchen hingegen nur ein einziges Mal auf. Übertragen auf die Sprache des Belegarchivs enthält das elektronische Belegarchiv nach der Aufnahme eines einzigen Buchs bereits knapp 12000 Stichwörter mit über 50.000 virtuellen Belegen.
Nachgeschlagen werden die Belege durch die Eingabe des Stichworts in eine Suchmaschine. Die Ergebnisse werden von der Suchmaschine in Form des Stichworts und einer kurzen Umgebung (KWIC-Zeile) mit einer bibliographischen Referenz zurückgegeben. Beispielsweise würde man für das Wort *Antichambrierer* finden:

```
die Agenten, die Reisenden, die _Antichambrierer_ dachten: Wat für 'ne Wolke von Weib
```
Koeppen, Wolfgang, Das Treibhaus, in: ders., Drei Romane, Frankfurt a.M.: Suhrkamp 1972 [1953], Seite 336

Abb. ... KWIC-Zeile mit bibliographischer Referenz für Antichambrierer

Im Unterschied zu einem traditionellen Beleg, bei dem der Kontext nach der Exzerption festgelegt ist, lassen sich in einem elektronischen Belegarchiv die Kontexte theoretisch beliebig erweitern. Bei Bedarf kann man sich so den ganzen Satz, die Sätze davor oder danach, den gesamten Absatz oder sogar das ganze Werk anzeigen lassen. Darüber hinaus kann man stets auf die Frequenzangaben innerhalb des Korpus Bezug nehmen, wodurch sich – je nach Qualität des Korpus – die Möglichkeit ergibt, Aussagen über die Gebräuchlichkeit des Wortes oder der Wortverbindung zu machen. Schließlich lassen sich Wortbildungsmuster auch noch nachträglich, d.h. nach der Exzerption ermitteln. Beispielsweise könnte der

Lexikograph bei Beschreibung des Eintrags *antichambrieren* die Häufigkeit der Belege und deren Verwendung für *Antichambrist* und *Antichambrierer* in allen Werken recherchieren, ohne dass diese vorher eigens für diesen Zweck hätten exzerpiert werden müssen.

Diese Flexibilität hat jedoch auch Nachteile. Für ein Buch mag es noch mit etwas Fleiß möglich sein, die interessanten von den uninteressanten Stichwörtern zu trennen: So würde man beispielsweise Eigennamen wie *Keetenheuve* oder redundante Belege von *und* per Hand aussortieren. Es ist jedoch unmittelbar einleuchtend, dass die manuelle Belegfilterung bei wachsenden Textmengen schnell an ihre Grenzen gelangt. Als minimale Größe für ein Textkorpus, welches als Korrektiv für Wörterbücher von Bedeutung ist, wird das 1993 erstellte British National Corpus (BNC) angesehen. Dieses umfasst etwa 100 Millionen aneinandergereihte Wortformen und somit potentiell 100 Millionen Belege. Eine so große Belegmenge kann nicht mehr vollständig per Hand nach neuen Wörtern, geschweige denn nach interessanten Verwendungsweisen von Wörtern durchgesehen werden. Auf der anderen Seite reduziert sich die astronomisch groß anmutende Zahl von 100 Millionen sehr schnell auf eine handhabbare Größe, wenn man diese Zahl in eine Anzahl von Büchern umrechnet. Legt man als Musterbuch Wolfgang Koeppens *Treibhaus* mit seinen rund 50.000 Textwörtern zugrunde, so sieht man, dass die Zahl von 100 Millionen aneinandergereihten Wörtern umgerechnet nicht mehr als 2000 Büchern im Umfang von Wolfgang Koeppens Treibhaus entspricht. Dies ist im Vergleich zu einem Belegarchiv für ein großes Wörterbuchvorhaben nicht sehr groß, wenn man bedenkt, dass für die Neubearbeitung der Buchstaben A-C des Wörterbuchs von Jacob Grimm und Wilhelm Grimm etwa 10.000 Quellen systematisch exzerpiert wurden. Man hat daher schon vor mehr als zehn Jahren in England, Frankreich und Deutschland begonnen, noch größere elektronische Korpora zu erstellen. Derzeit umfassen die hauseigenen Korpora der Verlage in der Regel wenigstens 500 Millionen aneinandergereihte Textwörter, für wissenschaftliche Zwecke genutzte Korpora gehen sogar darüber hinaus. So hat das DWDS-Ergänzungskorpus beispielsweise mehr als eine Milliarde Textwörter[2]. Rechnet man die Textmenge des Milliardenkorpus in Bücher um, so gelangt man zu einer Zahl von knapp 20.000 Bänden vom Umfang des „Treibhaus“. Nehmen wir an, dass eine Person ein Buch pro Woche vollständig und immer auf die gleiche Weise exzerpieren könnte. Dann entspricht dieser Umfang einer Exzerptionsarbeit von mehr als 250 Jahren. In diesem Korpus kommt alleine das Wort *und* 20,8 Millionen Mal vor, das Verb *gehen* 1 Million Mal, *leisten* noch 100.000 Mal, das Nomen *Teufel* noch knapp 25.000 Mal etc. Diese Belegzahlen dürften auch die letzten Zweifel nach der manuellen Auswertbarkeit zerstreuen.

Zwar ist damit gezeigt, dass mit Hilfe der vollständigen Exzerptionsmethode riesige Mengen von elektronischen Texten digitalisiert werden können, die mit manueller Arbeit wenn überhaupt, dann nur mit sehr großem Aufwand zu gewinnen wären. Ob und in welchen Bereichen diese großen Textkorpora die tatsächliche lexikographische Arbeit verbessern, hängt jedoch von zwei Aspekten ab, auf die im folgenden näher eingegangen wird: zum einen muss gezeigt werden, dass die riesigen Belegmengen automatisch so gut vorsortiert werden können, sie die Arbeit des Lexikographen tatsächlich unterstützen. Zum anderen muss gezeigt werden, dass die Korpora auch qualitative Vorteile gegenüber den traditionellen Belegsammlungen besitzen.

3. Linguistische Filter

[2] Das am IDS-Mannheim beheimatete interne IDS-Korpus umfasst sogar knapp 2 Milliarden aneinandergereihte Textwörter.

Die Tatsache, dass elektronische Suchsysteme in der Lage sind, jedes Wort zu indizieren und damit als Beleg zu inventarisieren, hat, wie in vorigen Abschnitt gezeigt, auf der einen Seite den Vorteil, dass dem Lexikographen kein „Wort durch die Lappen gehen“ kann. Auf der anderen Seite hat ein solches Verfahren, welches jedes Wort indiziert, den großen Nachteil, dass es Wichtiges von Unwichtigem nicht unterscheidet. Denn man wird wohl schwerlich behaupten können, dass alle von einem elektronischen System automatisch inventarisierten 20,8 Millionen Belege des Wortes „und“ im erweiterten DWDS-Korpus für den Lexikographen ebenso wichtig sind wie beispielsweise die Erfassung des oben erwähnten *Ceranfeldes*. Ist nun der Lexikograph bei der Auswertung der Textkorpora wieder zu einem Großteil auf seine eigene Sprachkompetenz zurückgeworfen und übernimmt damit selbst einen Teil der Belegexzerption, oder können die umfangreichen Belegmengen mittels computer-linguistischer Verfahren soweit vorsortiert werden, dass der Lexikograph nur noch die für ihn „interessanten“ Belege gezeigt bekommt?

Die in der Computerlinguistik verwendeten Techniken zur Belegextraktion und –sortierung nehmen einen breiten Raum in der Literatur ein und können hier daher im einzelnen nicht dargestellt werden; zudem ist dies für die Argumentation im Weiteren auch nicht notwendig. Konzeptuell gesprochen konzentrieren sich alle verwendeten Techniken auf drei verschiedene Bereiche: klassische Suchtechniken mit regulären Ausdrücken und Booleschen Operatoren, die linguistische Annotation sowie statistische Methoden.

Mit Hilfe regulärer Ausdrücke und Boolscher Operatoren lässt sich der Suchraum bereits für ein breites Spektrum von Suchanfragen genügend einschränken. Die in allen Suchmaschinen eingebaute Rechtstrunkierung lässt die Suche nach unterschiedlichen Endungen zu. Umgekehrt ermöglicht die in linguistischen Suchmaschinen implementierte Linkstrunkierung die Suche nach Präfixen. Ebenso kann man Platzhalter für einzelne Buchstaben setzen um nach Infixen zu suchen. Die Boolschen Operatoren UND, ODER, NICHT ermöglichen weitere Filterungen von Belegen. Alle diese Filter spielen sich jedoch auf der Zeichenebene ab.

Mit der linguistischen Annotierung verlässt man die Ebene der reinen Wortformensuche. Hier werden Wörter oder Wortverbindungen mit verschiedenen Informationen angereichert. Diese reichen von der Lemmatisierung, bei der die Wortform auf einen morphologische Stamm abgebildet wird, über die morphosyntaktische Annotierung, bei der jeder Wortform morphologische Eigenschaften und die lexikalische Kategorie zugeordnet wird, bis hin zur syntaktischen Annotierung, die im einfachsten Fall die Annotierung von Phrasen umfassen kann und bis zur vollen syntaktischen Analyse eines ganzen Satzes reicht (Heid u.a. 2000). Zusätzlich kann man die Wortformen mit semantischen Annotationen versehen, wie sie z.B. in WordNet (Fellbaum 1998) oder in seiner deutsche Variante GermaNet (Kunze 2000) vorliegen. Da bei der oben angesprochenen Korpusgröße keine Annotierung per Hand sinnvoll bzw. möglich ist und gleichzeitig automatische Methoden keine vollständige Abdeckung der Phänomene liefern, müssen entweder bei der Analysetiefe oder aber bei der Korrektheit bzw. Vollständigkeit der Analyse Abstriche gemacht werden. Dies beginnt bereits bei der am einfachsten erscheinenden Lemmatisierung. Bei sehr großen Korpora gibt es immer wieder Wortformen – beispielsweise Eigennamen –, die den morphologischen Analyseprogrammen unbekannt sind.

Die dritte Möglichkeit der Vorsortierung bildet die statistische Analyse (Manning & Schütze 1999). Im Wesentlichen geht es bei dieser zunehmend weiter verbreitete Analysemethode darum, statistische Auffälligkeiten zwischen zwei- oder mehr Wortformen zu berechnen und für den Lexikographen so darzustellen, dass die Belegdurchsicht erleichtert wird. Angewandt

wurden diese Methoden beispielsweise für die Extraktion von Adjektiv-Nomen Kollokationen (Church & Hanks 1990) oder Funktionsverbgefügen (Krenn 2000).

Im allgemeinen werden für die Korpusabfrage alle drei Möglichkeiten kombiniert. Eine ganze Reihe von Korpuswerkzeugen[3] hat sich mittlerweile etabliert, mit Hilfe derer man Boolesche Suchabfragen mit linguistischen Suchabfragen verknüpfen kann und ebenso statistische Abfragen stellen kann. Ein Bildschirmauszug des DWDS-Korpus illustriert die Möglichkeiten, die die Verknüpfung von Boolscher Suche und morpho-syntaktischer Annotation bietet. In dem Beispiel wird nach allen Adjektiven, die auf *–bar* enden, gesucht. Die so formulierte Suchabfrage filtert einerseits alle Nomen heraus, die auf *–bar* enden, andererseits findet sie auch die großgeschriebenen Adjektive am Satzanfang. Beides zusammen ließe sich nicht mit einem einfachen regulären Ausdruck erreichen. Natürlich setzt so eine Abfrage ein reiches Morphologiesystem voraus, welches in der Lage ist, Adjektive in Komposita und in derivierten Formen zu erkennen.

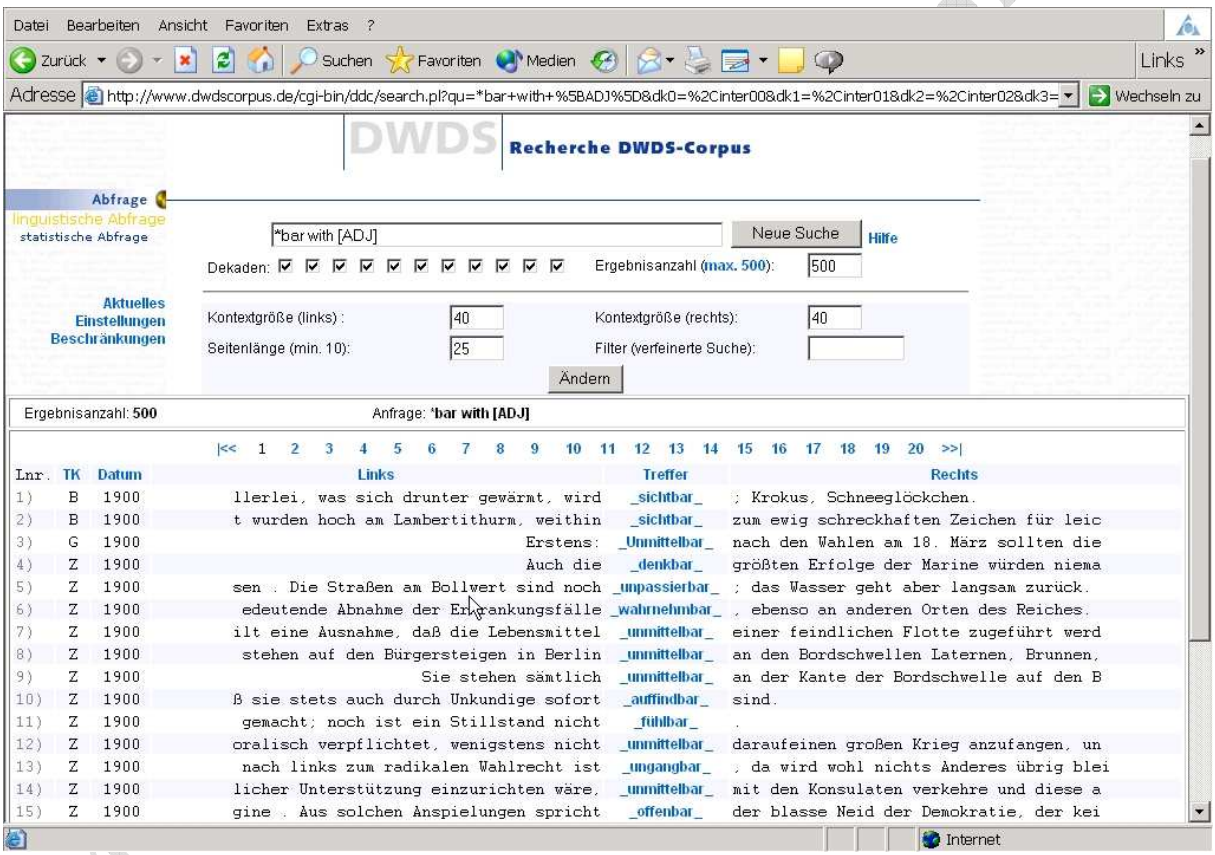

Abb. 1: Abfrage im DWDS nach Adjektiven, die auf *bar enden

Das im deutschen Sprachraum vermutlich am weitesten fortgeschrittene Werkzeug zur Neubearbeitung von Wörterbüchern stellt das System LexiView dar, welches in Zusammenarbeit mit der Universität Stuttgart und den Verlagen Duden-Langenscheidt entwickelt wurde (Heid 2000, 2004). Der in diesem Projekt entwickelte lexikographische Arbeitsplatz verbindet die drei oben angesprochenen Suchmöglichkeiten mit dem neuzubearbeitenden Wörterbuch in der Weise, dass es die Wortkandidaten für eine Neuaufnahme in das Wörterbuch oder aber auch die Herausnahme aus dem Wörterbuch

[3] wie SARA für das BNC, der in Stuttgart entwickelt Corpus Query Processor (CQP) oder das für das DWDS entwickelte DDC (Dialing DWDS Concordancer).

aufgrund von Frequenzinformationen automatisch vorberechnet und sie dem Lexikographen auf einer Oberfläche präsentiert. Die Korpusanalyse erfolgt auf drei Ebenen: Morpho-syntaktische Angaben der Wörterbücher werden mit den Fundstellen der Korpora abgeglichen, für Valenzangaben werden Beispielssätze extrahiert, und statistische Methoden werden verwendet, um die Angaben zu phraseologischen Verbindungen in Wörterbüchern zu ergänzen.

Einen etwas anderen Ansatz verfolgt der in Brighton entwickelte WordSketcher (Kilgarriff 2002). Hier werden zu einer Wortform syntagmatische Beziehungen mittels eines Parsers vorberechnet, statistisch verarbeitet und für den Nutzer so präsentiert, dass diesem alle statistisch auffälligen syntagmatischen Beziehungen in sehr kompakter Form auf einer einzigen Seite vorliegen. Beispielsweise würde das System für ein transitives Verb alle statistisch auffälligen Subjekte (genauer die Köpfe der Nominalphrasen, die Subjektkandidaten sind), direkten Objekte bzw. Präpositionalobjekte geordnet nach der regierenden Präposition präsentieren. Durch Hypertextverweise kann sich der Nutzer detailliertere Informationen zu all diesen Informationen einschließlich der Verknüpfung zu den Korpusbelegen machen.

4. Korpuserstellung

Das Verfahren, klassische Belegsammlungen durch die Aufnahme von großen Mengen gedruckter Texte zu ersetzen, erscheint sehr vielversprechend, da gedruckte Texte in zunehmendem Maße digital vorliegen und linguistischen Filter zunehmend besser in der Lage scheinen, die lexikographisch relevanten Informationen aus den riesigen Belegmengen zu extrahieren, womit der gegenüber den Zettelkästen gravierendste Nachteil der „blinden“ Belegsammlung kompensiert scheint. Zwar stellt es in der Tat kein großes Problem mehr dar, mehrere Zeitungsjahrgänge mit einem Umfang von mehreren Hunderttausend, ja sogar Millionen Artikeln zu sammeln und durchsuchbar zu machen. Repräsentativ für die Sprache ist eine solche Textmenge aber keinesfalls; vielmehr besteht das Risiko, dass solche teilweise wahllos zusammengestellten Textmengen Phänomene enthalten, die keinesfalls typisch für die Sprache im allgemeinen sind und somit die Sprachbeschreibung verfälschen. Welche Textsorten oder Genres sollte daher ein elektronisches Textkorpus enthalten und wie groß sollte es sein, um für die Neubearbeitung von Wörterbüchern hilfreich zu sein?
Das erste elektronische Korpus, das Brown-Corpus (Kučera und Francis 1967), formulierte dabei folgende Desiderata: das Korpus soll Textausschnitte aus möglichst ausgewogenen Genres enthalten, die Texte innerhalb einer Textsorte sollen durch Zufallsauswahl bestimmt werden und die vorher festgelegten Größenverhältnisse der Textsorten untereinander reflektieren, ferner sollten alle Texte aus möglichst einem Jahrgang bestehen. Das Brown-Corpus selbst besteht aus Texten (allesamt aus dem Jahre 1964) im Umfang von 1 Million aneinandergereihten Wörtern aus 15 verschiedenen Textsorten.

Genre	Anzahl von Textsamples zu je 2000 Wörtern
Presse: Reportagen	44
Presse: Kommentare	27
Presse: Rezensionen	17
Religion	17
Handwerk, Handel und Freizeit	36
Trivialliteratur	48
Lieratur, Biographien, Essays	75
Regierungskodumente	30

Geistes- und naturwiss. Schriften	80
Erzählungen allgemein	29
Kriminalromane	24
Science-fiction	6
Abenteuer- und Wildwestgeschichten	29
Liebesromane	29
Humor	9

Abb. 2: Zusammensetzung des Brown-Corpus

In der Literatur über Korpuslinguistik besteht seit langem Einigkeit darüber, dass das Idealziel der repräsentativen Abbildung der Sprache nicht zu erreichen ist (Rieger 1979), da mit empirischen Methoden nicht ermittelt werden kann, wie die Grundgesamtheit der Sprache beschaffen ist. Insbesondere ist die Unterteilung der Sprache in Textsorten umstritten, da weder über deren Anzahl noch über deren Gewichtung untereinander Klarheit herrscht (Ostdijk 1988, Bergenholtz 1989, Biber 1994). Man hat den Begriff der Repräsentativität daher durch denjenigen der Ausgewogenheit bezüglich der Textsorten ersetzt und gefordert, dass Textkorpora für die intendierten Zwecke hinreichend groß sein müssten. Da mit der Erstellung von Korpora die unterschiedlichsten Zwecke verfolgt werden können – Sprachentwicklung, Vergleich Mündlichkeit-Schriftlichkeit, Untersuchung von Regional- oder Jugendsprache, Neubearbeitung von einsprachigen Wörterbüchern –, folgt daraus, dass Korpora in Größe und Zusammenstellung stark variieren. Das Brown-Corpus und dessen deutsches Pendant, das Limas-Korpus, sind mit einem Umfang von einer Million laufender Textwörter sicherlich nicht geeignet, um die Neubearbeitung eines großen einsprachigen Wörterbuchs ausreichend zu unterstützen. Als minimale Größe hierfür gilt das etwa 100 Mal größere British National Corpus (BNC). In Ermangelung großer ausgewogener Korpora im deutschen Sprachraum wird für die Neubearbeitung von Wörterbüchern bislang vorwiegend auf Zeitungsquellen zurückgegriffen (Heid 2004): „Corpus material has been taken from freely available or specifically licensed newspaper texts, among others FR, Stuttgarter Zeitung 1992/93, a total of 350 m words. Our corpus is not balanced, as a general balanced corpus of German is only being created, e.g. at BBAW". Dieses Korpus, das im Rahmen des Projekts „Digitales Wörterbuch" an der BBAW erstellt wurde, soll nun kurz erläutert werden (vgl. dazu auch den Aufsatz von Wolfgang Klein in diesem Heft).

5. Das DWDS-Korpus

Das DWDS-Korpus umfasst insgesamt mehr als eine Milliarde laufende Textwörter und besteht aus einem Kernkorpus und einem Ergänzungskorpus. Das Kernkorpus, welches sich über das gesamte 20. Jahrhundert erstreckt, umfasst etwa 100 Millionen Textwörter; dies entspricht in etwa einer kleinen Bibliothek von ca. 1.500 Monographien (in Form von 120.000 Dokumenten). Etwa 40% davon wurde in mehr als zweieinhalb Jahren Arbeit mit bis zu 20 studentischen Mitarbeitern digitalisiert, der Rest wurde von Verlagen gekauft bzw. von Textgebern eingeworben. Die Texte des Kernkorpus sind gleichmäßig über das gesamte 20. Jahrhundert gestreut, die verschiedenen Fachsprachen und Textsorten sind angemessen repräsentiert. 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-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 vier anderen Bereiche setzen sich nach folgenden Auswahlkriterien zusammen:
Schöne Literatur (27%): darunter wird nicht nur die "hohe Literatur" verstanden, sondern auch die Unterhaltungsliteratur, die bislang lexikographisch 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 (26%): Diese Textsorte umfasst sowohl die überregionalen Tages- und Wochenzeitungen gedacht, aber auch einige regionale Blätter, die unter lexikographischen Aspekten oft besonders interessant sind; weiterhin an 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 Corpus 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 (22%): 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 (20%): 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".

Das DWDS hat mit elf Verlagen (*Aufbau, Diogenes Verlag, Eichborn, S. Fischer Verlagsgruppe, Hoffmann & Campe, Kiepenheuer & Witsch, K.G. Saur Verlag, Spiegel, Suhrkamp, Ullstein, ZEIT*), sowie mehreren öffentlichen (z.B. *Deutsches Rundfunkarchiv*) und privaten Textgebern (z.B. *Digitale Bibliothek*) Nutzungsvereinbarungen über rechtebehaftete Texte abgeschlossen und kann z.B. Werke von Thomas und Heinrich Mann, Martin Walser, Heinrich Böll, Jürgen Habermas oder Victor Klemperer für die Internetrecherchen zur Verfügung stellen (www.dwdscorpus.de). Schrittweise sollen die Nutzungsvereinbarungen für die Nutzung sämtlicher Texte des Kernkorpus abgeschlossen

werden. Eine bibliographische Datenbank[4] gibt einen Überblick über alle für das Kernkorpus digitalisierten Quellen sowie über den Stand der Nutzungsvereinbarungen:

Das Ergänzungscorpus umfasst über 900 Millionen Textwörter. Es ist ein sogenanntes „opportunistisches“ Corpus und besteht im wesentlichen aus Zeitungsquellen der 80er und 90er Jahre des 20. Jahrhunderts. Opportunistisch bedeutet jedoch nicht, dass es wahllos erstellt wurde. Alle Quellen sind bibliographisch referenzierbar, und bei der Aufbereitung wurde auf inhaltliche und qualitative Streuung geachtet. Neben *Frankfurter Allgemeine Zeitung*, *Neue Zürcher Zeitung* und *Süddeutscher Zeitung* wurden auch *Bild* oder *B.Z.* aufgenommen, neben *Die Zeit* und *Spiegel* sind auch *Konkret* oder die *tageszeitung* enthalten.

6. Rolle der Korpusqualität und -quantität

In diesem Abschnitt soll gezeigt werden, wie man je nach verwendetem Korpus zu ganz unterschiedlichen Ergebnissen kommen kann. Die Beispiele 1 und 2 illustrieren die Bedeutung von unterschiedlichen Textstreuungen von Korpora, die Beispiele 3 und 4 erläutern, in welchem Masse die Korpusgröße die Ergebnisse beeinflussen kann.

Beispiel 1: Korpora zur Aufdeckung von Archaismen
Im Projekt LexiView (Heid 2000) dient ein Zeitungskorpus im Umfang von 250 Millionen aneinandergereihten Textwörtern (tokens), bestehend aus taz, fr und dpa, zur Bewertung des deutschen Stichwortteils des Handwörterbuch Deutsch-Englisch von Langenscheidt: „....The use of corpus tools not only showed us words to be taken up into the new Großwörterbuch, it also showed dictionary corpses, old or rare words which miraculously had survived generations of lexicographers and revisions: *Immobilienmagnat* or *immobilisieren* are such words which used to be entries in the Handwörterbuch but had zero-evidence in the corpus." (Heid 2000:192f.). Auch wenn das primäre Ziel dieser Veröffentlichung darin bestand, den Nutzen computerlinguistischer Werkzeugen für die Neubearbeitung von Wörterbüchern zu zeigen, so ist die Schlussfolgerung an dieser Stelle voreilig. Legt man beispielsweise das erweiterte DWDS-Korpus zugrunde, so findet man für das Suchwort *Immobilienmagnat* 27 Treffer verteilt über verschiedene Quellen: FAZ, NZZ, SZ und taz. Die Suchanfrage *immobilisieren* liefert sogar 75 Treffer in verschiedenen syntaktischen Umgebungen: 21 Mal als Verb im Infinitiv, als Partizip oder als deverbales Adjektiv (z.B. *immobilisierte Banken*). Die Belege (1) und (2) illustrieren, dass die Verwendung dieser beiden Wortformen durchaus geläufig ist:

Beleg (1)
... ob der 58 Jahre alte Immobilienmagnat und Erbe des Eisenbahnnetzes des "Seibu"-Konzerns...
o.A., Heimlicher "Kaiser" des japanischen Sports Yoshiaki Tsutsumi häuft Geld wie Schnee, in: Die F.A.Z. auf CD-ROM (Jahrgang 1993), Frankfurt a.M.: Frankfurter Allgemeine Zeitung GmbH 1998 [1993]

Beleg (2)
Gemeinsam ist all diesen Plänen das Ziel, den Gegner zu immobilisieren, vorübergehend unschädlich zu machen.
Jean Guisnel, o.T. [In den Militärlabors ...], in: DIE ZEIT 09.02.1996, S. 34

Beispiel 2: Genusvariation bei Anglizismen

[4] www.dwdscorpus.de/cgi-bin/autoren/autorensuche

Korpora werden in Studien immer wieder zur Evaluierung von morpho-syntaktischen Angaben von Wörterbüchern herangezogen. Insbesondere können Frequenzangaben in Korpora Evidenz für morpho-syntaktische Angaben bilden. Dies bietet sich bei konkurrierenden Formen ebenso an wie bei unterschiedlicher Bewertung des Genus. In einer Studienarbeit[5] wurde die Genusverteilung einiger Anglizismen wie Blackout, Fitness oder Toast im DWDS-Kernkorpus, im IDS-Korpus und in Google verglichen. Weder die Größe der Korpora, wenn man die Texte im World Wide Web überhaupt so bezeichnen kann, noch deren Zusammensetzung sind vergleichbar: So stellt das DWDS-Kernkorpus, wie in Abschnitt 5 erläutert wurde, eine ausgewogene Auswahl über mehrere Textsorten dar, umfasst aber lediglich 100 Millionen aneinandergereihte Textwörter. Der in der Studienarbeit herangezogene Teil des IDS-Korpus umfasst 1,5 Milliarden Textwörter, enthält aber fast nur Zeitungstexte. Der von Google indizierte Teil des Internets umfasst mehr als 1 Milliarde Webseiten (URL). Es ist jedoch nahezu unmöglich, über deren Verteilung und Qualität genauere Aussagen zu treffen (Rundell 2000). Ergänzend zu dieser Studie habe ich auch noch das DWDS-Ergänzungskorpus (1 Milliarde Textwörter) herangezogen.
Für das zu betrachtende Nomen *Blackout* erteilen die großen einsprachigen Gegenwartswörterbücher unterschiedliche Auskünfte. So geben Wahrig[6] und das Wörterbuch der deutschen Gegenwartssprache (WDG) [7] nur das Neutrum an, wohingegen der 10-bändige Duden (1999) sowohl Maskulin als auch Neutrum für möglich hält. In den vier Korpora erhält man zu dieser Frage völlig unterschiedliche Ergebnisse (s. Abbildung 3). Einig sind sie sich nur darin, dass die Mehrzahl der Belege im *Blackout* im Maskulin enthalten. Am knappsten fällt die Entscheidung bei Google[8] und im IDS-Korpus aus. Legt man nur diese beiden Korpora zugrunde, so würde man daraus folgern, dass die Verwendung von *Blackout* im Neutrum durchaus geläufig ist. Diese Schlussfolgerung wäre hingegen in den beiden DWDS-Korpora ganz anders. Im DWDS-Kernkorpus ist kein einziger der 14 *Blackout* Belege eindeutig dem Neutrum zuzuordnen; auch im DWDS-Ergänzungskorpus lassen sich nur drei Belege im Neutrum finden. Die Ergebnisse in den DWDS-Korpora weisen somit auf eine eindeutige Präferenz des Maskulins hin. Ein zweiter Blick auf die Trefferliste zeigt darüber hinaus, dass die Belege mit Neutrum der Jugendsprache oder zumindest einem anderen Register zuzuordnen sind als die Belege im Maskulin. Dies illustrieren die beiden folgenden Belege:

Beleg (3)
Bremen (taz) - Der Streit um das Blackout eines Radio-Bremen-Moderators ist noch nicht zu Ende.
Bremer Blackouts, in: die tageszeitung - 12 ½ Jahre taz auf CD-ROM, Berlin: Contrapress-Media-GmbH 1999 [1998]

Beleg (4)
Die einzige Erzählung, die ein Happy-End hat, ist die erste, "Bankraub", und auch da darf man sicher sein, daß nur der Blackout es dem Erzähler erläßt, von den Qualen des genossenen Glücks zu berichten.
Zeitzünder in der Couchecke, in: F.A.Z.-Buchkritik '98, Frankfurt a.M.: Frankfurter Allgemeine Zeitung GmbH 1998 [1998]

Schließlich weicht auch die relative Beleganzahl in den jeweiligen Korpora voneinander ab: das 1,5 Milliarden Wörter umfassende IdS-Korpus enthält 54 Treffer, wohingegen das nur 100 Millionen umfassende DWDS-Ergänzungskorpus 86 Treffer enthält.

[5] www2.hu-berlin.de/korpling/lehre/ws-2003/hs-phaenomene-deutsch/Phaenomene-Anglizismen-klein.pdf
[6] www.wissen.de
[7] http://www.dwds.de/wdg,
[8] Abfrage vom 11.1.2004

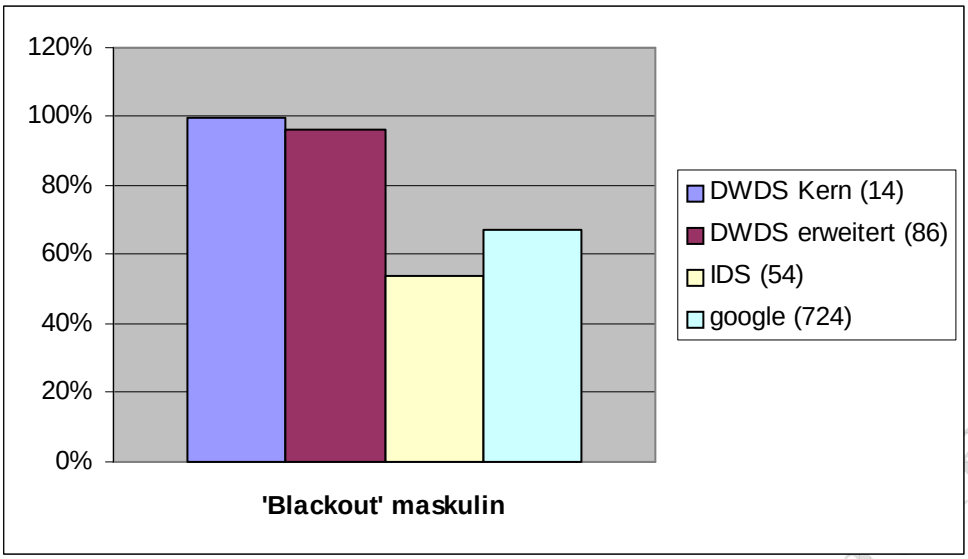

Abbildung 3: Blackout

Der Vergleich der drei Korpora zeigt, dass Korpusqualität und –quantität für die Ergebnisse eine sehr große Rolle spielen. In Übereinstimmung mit anderen Studien (z.B. Dumais 2002) weisen diese Ergebnisse darauf hin, dass die Verbesserung der Quantität und Qualität der Daten eine mindestens ebenso hohe Bedeutung zuzuschreiben ist wie der Verbesserung von Extraktionsmethoden.

Beispiel 3: Variation von statistisch signifikanten Kollokationen bei wachsendem Korpus

Der Vergleich des DWDS-Kernkorpus mit dem DWDS-Ergänzungskorpus zeigt, dass statistische Untersuchungen je nach dem zugrundeliegenden Korpus, zu ganz unterschiedlichen Ergebnissen führen können. Beispielsweise erhält man für das Verb *hegen* in seiner Bedeutung als Nominalisierungsverb im Sinne von *Hoffnung hegen* oder *Zweifel hegen* unter Verwendung des gleichen statistischen Maßes bedeutend mehr prädikative Nomen im DWDS-Ergänzungskorpus als im DWDS-Kernkorpus. Die Auswertung des DWDS-Kernkorpus liefert in abnehmender Signifikanz gemäß Log-Likelihood (Dunning 1993) folgende signifikante Kollokationspartner:

> *Hoffnung, Zweifel, Wunsch, Erwartungen, Befürchtungen, Argwohn, Verdacht, Bedenken, Vorurteile, Hoffnungen, Vertrauen, Glauben, Gedanken* (13)

Demgegenüber ergibt die Auswertung des erweiterten DWDS-Ergänzungskorpus eine Liste von 23 verschiedenen statistisch signifikanten prädikativen Nomen:

> *Zweifel, Hoffnung, Verdacht, Befürchtungen, Erwartungen, Hoffnungen, Wunsch, Bedenken, Ambitionen, Absichten, Befürchtung, Groll, Pläne, Argwohn, Absicht, Erwartung, Vorurteile, Mißtrauen, Wünsche, Illusion, Sympathien, Vermutung, Vertrauen.*

Beispiel 4 Untersuchung von Idiomen

Von Idiomen weiß man, dass sie im allgemeinen in Texten vergleichsweise selten vorkommen, obwohl sie im Bewusstsein der Sprecher in der Regel gut verankert sind. Vor

dem Hintergrund der oben gemachten Feststellung, dass die Neubearbeitung von vier großen englischen Wörterbüchern (mit) auf der Grundlage des BNC gemacht wurden (Rundell 1996), und das BNC somit als Korpus minimaler Größe für lexikographische Untersuchungen gilt, stellt sich die Frage, ob die Größe von 100 Millionen aneinandergereihter Textwörter für die Bearbeitung von Idiomen ausreichend ist. Dieser Frage sind wir in einer Studie im Rahmen des Projekts „Kollokationen im Wörterbuch“ (s. den Beitrag von Christiane Fellbaum in diesem Heft) auf der Basis von 46 Verb-Nomen-Idiomen nachgegangen. Die untersuchten Idiome wurden auf der Basis des Duden Band 11 Idiome zufällig ausgewählt. Beispiele hierfür sind *sich eins ins Fäustchen lachen, Kohldampf schieben, einen in der Krone haben, sich mit etw./jdn in die Tinte setzen.* Im Rahmen der Idiombeschreibung des Kollokationenprojekts wurden alle Vorkommen der 46 Idiome im DWDS-Ergänzungskorpus per Hand ermittelt und syntaktisch beschrieben. Hieraus ergibt sich folgende Verteilung:

- zwischen 1 und 10 Vorkommen: 9 Idiome
- zwischen 11 und 25 Vorkommen: 13 Idiome
- zwischen 26 und 100 Vorkommen: 15 Idiome
- mehr als 100 Vorkommen: 9 Idiome

Darüber hinaus führten wir eine Studie durch, um die Verteilung von Idiomen im DWDS-Ergänzungskorpus zu beschreiben (Geyken 2004). Ziel dieser Arbeit war es, das „Vorkommenswachstum“ der Idiome beschreiben sollte, um daraus eine sinnvolle minimale Korpusgröße für die Studie von Idiomen abzuleiten. Hierzu wurde das DWDS-Ergänzungskorpus in 100 Teilkorpora aufgeteilt, wobei durch sampling-Prozeduren darauf geachtet wurde, dass die Verteilung jedes Teilkorpus der Verteilung der Gesamtkorpus entspricht. Durch Abgleich der automatisch ermittelten Treffer für jedes Idiom in jedem Teilkorpus mit den per Hand ermittelten Satzbelegen kann dann berechnet werden, wie oft jedes Idiom in jedem Teilkorpus vorkommt. Durch Vereinigung der Teilkorpora bzw. durch Summenbildung der Vorkommen für jedes Idiom erhält man eine Wachstumskurve, die in Abbildung 4 für vier ausgewählte Idiome dargestellt ist:

- *jmd. hat etw. mit der Muttermilch eingesogen/aufgesogen (171 Belege)*
- *jmd. trinkt (einen) über den Durst (127)*
- *jmd. schiebt Kohldampf (63)*
- *jmd. macht Schmu (15)*

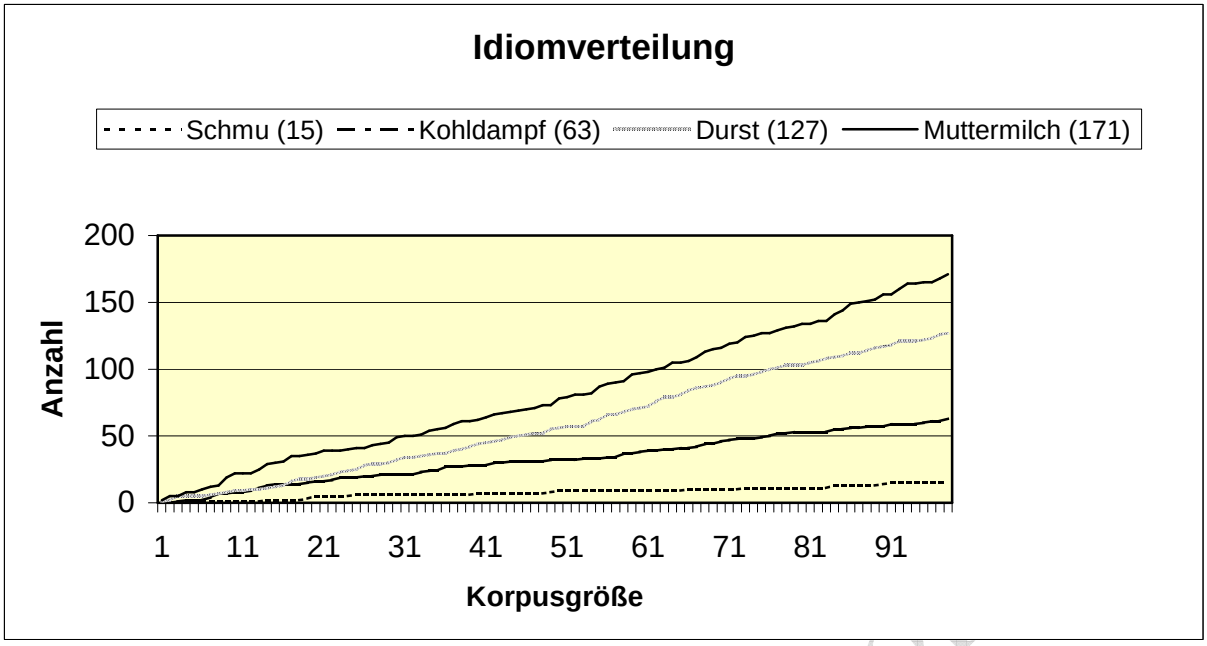

Abb. 4: Vorkommenswachstum der Idiome im DWDS-Ergänzungskorpus

Das gleichmäßige Wachstum bestätigt die Korrektheit der Sampling-Prozedur. Sie zeigt auch, dass ein Korpus von der Größe des BNC für eine Idiomuntersuchung nicht ausreicht. In der Abbildung entsprechen 100 Millionen aneinandergereihte Textwörter 10 Teilkorpora des Samplings, also dem Wert 10 auf der Abszisse. Hochfrequente Idiome hätten bei dieser Korpusgröße im Schnitt gerade einmal 20 Belege, bei den weniger frequenten wären es nur noch 4-10 Belege, bei den selteneren müsste man sich sogar im Schnitt mit einem Einzelbeleg begnügen. Eine Untersuchung von Abweichungen, die ja gerade bei Idiomen von Interesse ist, wäre somit gänzlich unmöglich.

7. Größe von Korpora: ein Vergleich mit Wörterbüchern

Die Größe von Korpora lässt sich sowohl anhand der aneinandergereihten Wörter (tokens) als auch nach der Zahl der verschiedenen Wörter (types) messen. In den Korpusstudien wird das Wort hier in einem naiven Sinne, nämlich als Zeichenkette zwischen zwei Leerstellen charakterisiert. Die Zahlen in Abbildung 5 zeigen zumindest zweierlei. Auf der einen Seite wird sichtbar, dass die Zahl der verschiedenen Textwörter im Deutschen wesentlich schneller wächst als in den Textkorpora der englischen Sprache. Dies lässt sich vergleichsweise einfach dadurch erklären, dass die Wortbildung und vor allem die Komposition im Deutschen produktiv sind, d.h. dass potentiell unendlich viele Wörter gebildet werden können. Auf der anderen Seite legen die Zahlen zur Anzahl der verschiedenen Wörter in den Korpora nahe, dass diese produktiven Mechanismen der Sprache auch genutzt werden, da die Anzahl der verschiedenen Wörter in den Korpora nicht gegen eine feste Obergrenze gehen, sondern mit zunehmender Textmenge wachsen. Zur Berechnung solcher Wortwachstumskurven liegen bereits Studien vor (Senellart 1996)[9].

[9] Eine entsprechende Studie zu einer Wortwachstumskurve auf der Grundlage der DWDS-Korpora ist in Vorbereitung.

Korpusname	laufende Textwörter (tokens)	verschiedene Textwörter
Brown Corpus	1 Million	50.000
Limas Corpus	1 Million	110.000
Brititsh National Corpus	100 Millionen	650.000
DWDS-Kernkorpus	100 Millionen	2,2 Millionen
erweitertes DWDS Korpus	1 Milliarde	9 Millionen

Abb. 5 [10]

Wörterbuch	Stichwörter
WDG	88.000
Duden	200.000
Grimm	333.000

Abb. 6

Der Frage nach dem Vergleich von Stichwortanzahl in Wörterbüchern und Korpusgröße soll im folgenden auf der Basis bereits bestehender Korpora nachgegangen werden. Vergleicht man die Zahlen in Abb. 5 und Abb. 6, 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 verschiedenen Wörtern 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 das Vierfache erhöht sich diese Zahl beim DWDS-Ergänzungskorpus. Dieses enthält etwas mehr als neun Millionen verschiedene Wortformen. Im Vergleich dazu mutet 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 330.000 Stichwörtern, klein an. Lässt sich hieraus ableiten, dass Korpora einen immensen Schatz an lexikographischem Material enthalten, der für die Lexikographie gehoben werden muss?
Zwar ist dieser Gedanke reizvoll, jedoch ist die Argumentation über die Anzahl der verschiedenen Wortformen alleine nicht ausreichend, da die Stichwortanzahl in Wörterbüchern etwas anderes misst als die Anzahl der verschiedenen „Wörter“ in Korpora. Insbesondere enthalten diese nach der oben erwähnten Zählweise auch Kardinalzahlen, Typenbezeichnungen oder Eigennamen. Auch fremdsprachliche Wortformen werden enthalten sein, da Bücher, Zeitschriften und Zeitungen zuweilen fremdsprachige Zitate enthalten. Der weitaus größte Anteil des lexikographisch uninteressantem Material, so könnte man als Kritiker des Zahlenarguments fortfahren, besteht aber in der großen Menge transparenter Komposita. Beispielsweise enthält das DWDS-Kernkorpus zu *Tür* die transparenten und damit lexikographisch uninteressanten *Holztür, Stahltür, Badtür, Schlafkammertür, Schlafzimmertür, Stalltür, Stubentür, Wohnzimmertür, Wohnungstür, Zimmertür,* etc. Es stellt sich somit die Frage, ob Korpora nach Abzug all dieser Wortformen noch lexikographisch interessantes Material enthalten. Der Nachweis darüber, dass sehr große Korpora auch Wörterbuchlücken im großen Maßstab enthalten, besteht in der Aufgabe nachzuweisen, wie die unbekannten und gleichermaßen lexikographisch interessanten Wörter aus diesen riesigen Textmengen effektiv extrahiert werden können.
Darüber hinaus wird immer wieder darauf verwiesen, dass auch sehr große Korpora nicht alle Stichwörter eines Wörterbuchs enthalten und somit in Vielfalt hinter den Belegarchiven der

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

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).
Vor dem Hintergrund der Tatsache, dass mittlerweile 10-fach größere Korpora als das von Hausser erwähnte BNC zur Verfügung stehen, stellt sich die Frage nach der Vollständigkeit von Korpora aufs Neue. Der Frage, inwieweit die beiden DWDS-Korpora, das Kern- und das Ergänzungskorpus als Korrektiv für große einsprachige Wörterbücher dienen können, wird im folgenden Abschnitt nachgegangen.

8. Korpora: Wörterbuchergänzung und –korrektiv

Das in der Einleitung angeführte Zitat über das „durch die Lappen gerutschte" *Ceranfeld* kann den Anschein erwecken, als würden Textkorpora lediglich für die Erfassung von Neologismen Verwendung finden. Dass dem nicht so ist, davon zeugen einige Projekte, die elektronische Textkorpora für die Neubearbeitung von Wörterbüchern in nahezu allen Bereichen der Makro- und Mikrostruktur einsetzen (z.B. Heid 2000, Heid 2004, Rundell 1996). 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. 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).

Korpora zur Ergänzung von Beispielen

Besonders naheliegend ist die Nutzung von Korpora zur Ergänzung von Belegbeispielen, insbesondere bei Einträgen mit einer relativ kleinen Beleggrundlage. Beispielsweise verzeichnet die Neubearbeitung des "Deutschen Wörterbuchs" von Jacob Grimm und Wilhelm Grimm unter dem Stichwort *Angstkauf* (^{2}DWB, Bd.2):

ANGSTKAUF m. *großeinkauf von waren aus sorge, daß diese knapp werden könnten* (zu angst f. 1): 1925 als die kurz nach kriegsausbuch einsetzende spekulation in verbindung mit angstkäufen die preise .. in die höhe trieben gemeines 2,3, F. 1959 H.C. trägt wieder einen karton tafelbutter zum wagen. „es gibt .. noch zu viele .. menschen, die sich zu sogenannten angstkäufen hinreißen lassen .." *berl. ztg.* (27.11.)

In diesem Artikel wird der Erstbeleg mit 1925 angesetzt, der letzte Beleg des 1998 erschienen Bands wird 1959 datiert. Das DWDS-Korpus enthält 15 Belege, der erste davon aus dem Jahre 1916, bei dem es um Angstkäufe im Zuge der Zuckerknappheit geht. Der letzte Beleg im DWDS-Korpus ist von 1998. Somit erstreckt sich die Beleggrundlage für Angstkauf über

ein weit größeres Zeitintervall als das in der Neubearbeitung des Deutschen Wörterbuch angegebene. Darüber hinaus fehlt im Wörterbuchartikel, dass das Wort im Singular nicht gebräuchlich ist. Die diesbezüglich eindeutige Korpusbeleggrundlage – alle Belege enthalten den Plural von Angstkauf – legt jedoch nahe, eine solche Anmerkung vorzunehmen.

Korpora als Korrektiv für die Stichwörter am Beispiel des WDG

Das Wörterbuch der deutschen Gegenwartssprache (WDG) wurde in Berlin an der Deutschen Akademie der Wissenschaften (ab Oktober 1972: Akademie der Wissenschaften der DDR) zwischen 1952 und 1977 erarbeitet. Das WDG umfasst über 4.500 Seiten und enthält etwa 88.000 Stichwörter[11]. Das WDG versteht unter deutscher Gegenwartssprache "außer der so charakterisierten, heute geschriebenen und gesprochenen Sprache der bildungstragenden Schicht auch die Sprache der in unserer Zeit noch gelesenen, lebendigen deutschen Literatur der Vergangenheit" (Vorwort zu Bd. 1, S. 4). Deshalb zählten zum Quellenbestand, der der Ausarbeitung der Wörterbuchartikel zugrunde lag, nicht nur die wichtigsten Gegenwartsautoren des ganzen deutschen Sprachbereichs bis in die siebziger Jahre des 20. Jahrhunderts, sondern auch zahlreiche ältere Texte deutscher Autoren seit Lessing und Kant. Genaue Auskunft darüber gibt das alphabetische Verzeichnis der Quellen (Bd. 6, S. 4559 - 4579).
Die Debatte, inwieweit der Abgleich von Korpusbelegen mit der Stichwortliste ideologische Vorannahmen zutage fördert, soll hier nicht geführt werden. Vielmehr geht es darum zu zeigen, welche Aussagen man über die Stichwortliste des WDG mit einem sehr großen Korpus treffen kann. Der Abgleich der Stichwortliste mit dem DWDS-Ergänzungskorpus führt zu einer Liste von etwa 2000 Wörtern, die im Korpus höchstens 1 Mal als Stamm oder als flektierte Form vorkommen[12]. Bei der Durchsicht dieser Liste stößt man auf verschiedene Ursachen.

Abweichende Orthographie: Das WDG verzeichnet *Kapriccio*, *Koserie* oder *Monstrefilm*, die allesamt in dieser Graphie unüblich sind. Hingegen sind *Capriccio, Causerie* bzw. *Monsterfilm* 1182, 76 bzw. 55 Mal belegt.

Fugenformen: Das WDG führt bei einer ganzen Reihe von Stichwörtern orthographisch Varianten auf. Beispielsweise bedeutet die Schreibweise *Abenteu(r)erromantik*, dass das Wort sowohl mit als auch ohne „r“ geschrieben werden kann. Für die Variante mit „r“ findet man jedoch keinen einzigen Treffer im Korpus, wohingegen es für das Kompositum mit „r“ 16 Belegstellen gibt. Weitere Beispiele hierfür sind der *Abfahrt(s)hang* und die *Miet(s)preiserhöhung*. Ersteres taucht ohne das Fugen –s kein einziges Mal im Korpus auf, für letzteres hingegen lassen sich mit dem Fugen –s keine Belege finden, wohingegen sich für Form *Abfahrtshang* mehrere hundert Belegstellen finden lassen und damit überraschenderweise häufiger als die *Mietpreiserhöhung* (66 Mal) im DWDS-Korpus vorkommt.

Ungebräuchliche Derivationen: Die beiden vom Nomen *Moslem* abgeleiteten Adjektive *moslemisch* und *mosleminisch* werden im WDG gleichrangig aufgeführt, obwohl das Korpus für die erste Variante 3686 Treffer und für die zweite Variante nur einen einzigen Treffer verzeichnet. Ähnliches gilt für das Verb *picheln*, welches im WDG mit den

[11] Nimmt man diejenigen Komposita hinzu, die am Ende eines Wörterbuchartikels verzeichnet sind, so kommt man auf knapp 120.000 Stichwörter.
[12] Die genaue Zahl haben wir bislang noch nicht ermittelt, da das WDG einige orthographische Varianten wie beispielsweise Abenteurerurlaub oder Abfahrthang enthält, für die das im Projekt verwendete Morphologieprogramm noch nicht alle Vollformen erzeugen kann.

Wortbildungsmöglichkeiten der Nominalisierung *Pichelei* bzw. den Partikeln *aus-* und *weg-* in *auspicheln* und *wegpicheln* verzeichnet ist. Im Korpus taucht von diesen vier Formen nur das Basisverb *picheln* mit einer Frequenz von 36 auf.

Ungebräuchliche Komposita: In demselben semantischen Feld findet sich das Adjektiv *sternhagelbesoffen*. Hierfür findet sich folgender Eintrag:

> **stern-, -hagelbesoffen** /*Adj.*/ salopp derb *sinnlos betrunken*; **-hagelvoll** /*Adj.*/ salopp *vgl.* -hagelbesoffen: er war s.; s. schwankte er nach Hause

Die Frequenzanalyse im DWDS-Korpus ergibt, dass *sternhagelbesoffen* gar nicht, das durch Referenz darauf definierte *sternhagelvoll* hingegen vierzig Mal im DWDS-Korpus auftaucht. Darüber hinaus gibt es noch das dazu synonyme *sturzbesoffen*, welches im WDG gar nicht, im Korpus hingegen mit 43 Belegen am häufigsten von allen drei Wortformen verzeichnet ist. Weitere Beispiele hierfür sind die Adjektive *tailleneng*, *tizianblond*, *Aalwanderung*, *Abblasehahn* und die *Abortbrille*, die allesamt im DWDS-Ergänzungskorpus nicht vorkommen. Hingegen kommt das zum letzten Kompositum synonyme *Klobrille* 160 Mal vor, im WDG ist es jedoch nicht verzeichnet. Ein weiteres Beispiel ist die als DDR-Neuprägung benannte *Uranuhr*, welche im Korpus nicht verzeichnet ist. Auf der anderen Seite gibt es für das geläufige *Atomuhr* (185) im WDG keine Erklärung.

„Blinde Flecken" des Korpus

Umgekehrt weisen aber einige Wortformen des WDG auf „blinde Flecken" im Korpus hin. Beispielsweise ist der *Pomeranzenlikör* aus dem Backwesen nicht im Korpus belegt, auch die Kinderlieder und -literatur sind unterrepräsentiert – darauf weisen die im Korpus nicht vorhandenen WDG-Einträge *Puthenne* und *Heiabett* hin. Desweiteren scheint es im Korpus Lücken im Übergangsbereich von der Allgemein- zu Fachsprache bei den Berufssparten zu geben. Hier fehlt beispielsweise das *Nackenleder* aus der Feuerwehrsprache, ebenso fehlen gängige Begriffe aus der Stahlherstellung. Hier führt das WDG zwei Verfahren zur Stahlherstellung im Bodenblasverfahren an: das *Thomasverfahren* und das *Bessemerverfahren*. Beide kommen im Korpus kein einziges Mal vor. Auf der anderen Seite verzeichnet das WDG weder das *Rennfeuer* noch den Term *Bodenblasverfahren* selbst, obwohl diese im Korpus verzeichnet sind. Auch manche älteren Wortformen enthält das Korpus naturgemäß aufgrund seiner Ausrichtung auf das 20. Jahrhundert nicht. Dazu zählt beispielsweise das *Unschlittlicht*, welches im Gutenberg-DE[13] Archiv mehrfach vertreten ist.

Vollständigkeit der Stichwortlisten in Wörterbüchern

Ging es im vorigen Abschnitt um diejenigen Stichwörter, die im Korpus nicht oder zu selten belegt sind, so soll nun auf den Beitrag eingegangen werden, den Korpora leisten können, um zur Vollständigkeit der Wörterbücher beizutragen. Zwar machen automatische Verfahren zur Identifikation neuer Wortformen aus Korpora große Fortschritte – je größer die Korpora, desto bessere Ergebnisse können statistische Verfahren liefern. Automatische Verfahren können jedoch keinesfalls die lexikographischen Kompetenz ersetzen. Die Entscheidung darüber, ob es sich bei einer neuen Wortform um eine lexikalisierte Form oder lediglich eine transparente Bildung handelt, wird auf absehbare Zeit vom Lexikographen getroffen werden. 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- 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.

[13] vgl. http://gutenberg.spiegel.de

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. Diese Vorgehensweise wird im nächsten Abschnitt an einem größeren Beispiel illustriert. Hierfür werden alle Kompositionen von *Selbst* im Korpus mit denen des 10-bändigen Duden abgeglichen. Im zweiten Teil dieses Abschnitts gehen wir kurz auf die Schwierigkeit der Identifikation von Neologismen in Wörterbüchern ein.

Vollständigkeit der Komposita

Zur Illustration haben wir hier alle *Selbstkomposita* des 10-bändigen Duden (1999) mit den Komposita im Korpus verglichen. Der Duden verzeichnet 244 Einträge von *Selbstabholer* über *Selbstbedienung* und *Selbsterfahrung* bis hin zu *Selbstzweifel.*

Demgegenüber enthält das DWDS-Ergänzungskorpus 10934 verschiedene Selbstkomposita, die wir halbautomatisch auf 7180 verschiedenen Lemmata abgebildet haben. 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[14].

> *Selbstablieferer (119), Selbstaufklärung (128), Selbstauskunft (182), Selbstaussage (103), Selbstbedienungsmentalität (143), Selbstbefragung (312), Selbstbehauptungswille (139), Selbstbereicherung (123), Selbstbeschäftigung (105), Selbstbeschreibung (380), Selbstbewegung (116), Selbstbezüglichkeit (184), Selbstbindung (175), Selbstbiographie (250), Selbstblockade (212), Selbstdemontage (109), Selbstentblößung (194), Selbsterforschung (185), Selbsterlösung (105), Selbstermächtigung (100), Selbsterneuerung (111), Selbstfeier (122), Selbstfinanzierungsgrad (121), Selbstgefährdung (122), Selbstherrschaft (107), Selbsthilfeeinrichtung (137), Selbsthilfeprojekt (179), Selbsthilfezentrum (219), Selbstisolation (136), Selbstisolierung (165), Selbstlauf (162), Selbstmordanschlag (296), Selbstmordattentäter (801), Selbstmordattentat (260), Selbstmordgefahr (135), Selbstnutzer (108), Selbstoffenbarung (111), Selbstopfer (108), Selbstportrait (227), Selbstrechtfertigung (197), Selbstregierung (727), Selbstregulierung (450), Selbsttherapie (172), Selbstverpflichtung (1669), Selbstverständigung (284), Selbstverteidigungsgruppe (127), Selbstverteidigungskurs (119), Selbstverwaltungsrecht (134), Selbstverwaltungsorgan (131), Selbstverwaltungsbehörde (141), Selbstverwaltungsgremium (102), Selbstvorwurf (104), Selbstwert (171), Selbstwiderspruch (118), Selbstzitat (118)*

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*.

Umgekehrt gibt es Duden-Einträge, die im Korpus entweder gar nicht oder nur ganz selten vorkommen:

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

Selbstabholerin (0), Selbstanschuldigung (5), Selbstansteckung (4), Selbstanzeigerin(0), Selbstbeköstigung (5), Selbstbiografie (1), Selbstbucherin(0), Selbstentlader (0), Selbsterzeugerin(0), Selbstfahrerin (1) , Selbstinserent(0), Selbstladevorrichtung (0), Selbststellerin(0), Selbstverlegerin(0), Selbstverpflegerin (alle 0), Selbstverstand (0)

Vor allem die movierten Formen in dieser Liste erscheinen fragwürdig: zwar entsprechen sie den Wortbildungsregeln, die Tatsache, dass sie im Gegensatz zu ihren nicht-movierten Stammformen kein einziges Mal im Korpus vorkommen, spricht jedoch dafür, dass es sich bei diesen Einträgen um Wörterbuchartefakte handelt.

Gestützt wird die wichtige Rolle von Korpora bei der Neubearbeitung der Komposita in Wörterbüchern durch die Beobachtung, dass die Anzahl der verschiedenen Komposita mit zunehmender Korpusgröße wachsen. Wir haben dies im DWDS-Ergänzungskorpus exemplarisch an den Selbstkomposita überprüft. Hierzu haben das Korpus eingelesen und an sechs verschiedenen Messpunkten die Anzahl der verschiedenen Selbstkomposita ermittelt. Auch wenn sich daraus noch keine statistisch saubere Wachstumskurve ableiten lässt – man müsste hierzu ein statistisch valide Stichprobe erstellen -, so zeigen die Messpunkte eindeutig, dass die Anzahl der verschiedenen Selbstkomposita mit der wachsender Korpusgröße zunimmt.

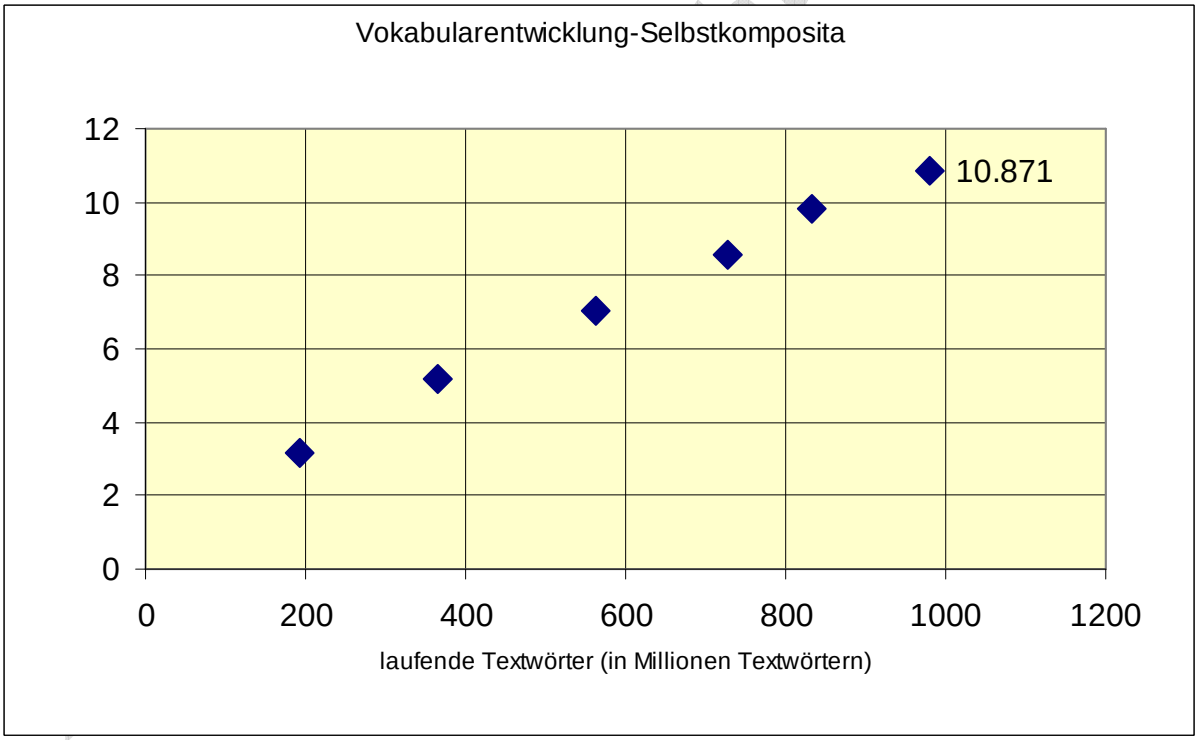

Abb. 7: Zunahme der Selbstkomposita im DWDS-Ergänzungskorpus

Identifikation von Neologismen

Neologismen, wenn sie sich nicht auf der Basis bereits bekannter Stichwörter suchen lassen, sind sehr viel schwieriger zu finden, da einerseits der Suchraum aufgrund der enormen Wortanzahlen in den Korpora zu groß ist, um ihn für die manuelle Analyse erschließbar zu machen, andererseits handfeste Kriterien für die Einschränkung des Suchraums Mangelware sind. Die folgenden Bemerkungen liefern hierfür keine Lösungen, sondern beschreiben lediglich mögliche Ansätze zur Einschränkung des Suchraums.

Der einfachste Filter zur Reduktion der großen Datenmengen stellt das Frequenzkriterium dar. Hierbei wird die Tatsache ausgenutzt, dass Textkorpora näherungsweise der Zipfschen Verteilung genügen (vgl. Manning & Schütze 1999:23). Dies impliziert insbesondere, dass nur sehr wenige Wortformen sehr oft, die meisten jedoch nur ein- oder zweimal vorkommen. Im DWDS-Ergänzungskorpus mit seinen 9.099.128 verschiedenen Wortformen findet sich dies bestätigt: 5.378.322 Wortformen kommen genau einmal, 1.183.751 zweimal, 532.415 Wortformen genau dreimal, 315535 genau viermal etc. Nur 1.036.590 verschiedene Wortformen kommen 10 Mal und mehr vor.
Mit Hilfe linguistischer Filter lässt sich diese Menge weiter einschränken. Eindeutig fremdsprachliches Material kann man herausfiltern, wenn dieses in längere fremdsprachliche Passagen eingebettet ist. Ferner lassen sich Eigennamenfilter einsetzen (Mikheev 1999), mit deren Hilfe sich vermeiden lässt, dass dem Lexikographen neue „interessant" aussehende Wortformen wie *Zumdick, Milchraum* (Sportler, vgl. Heid 2000), *Armatrading* (Musikerin), *Holzmann* (Unternehmen), oder *Hegermühle* (Ortsname) vorgelegt werden.

Korpora als Grundlage für rückläufige Wörterbücher

Rückläufige Wörterbücher ordnen der Wortschatz umgekehrt alphabetisch an. Sie beginnen mit Wörtern, die auf ‚a' enden, und enden mit Wörtern, die auf ‚z' enden. Mit dieser Anordnung werden Wörter gruppiert, die gleich oder ähnlich enden, z.B. *hasten, kasten, lasten, rasten, tasten*. Rückläufige Wörterbücher finden ihre Verwendung beispielsweise als Reimwörterbuch oder als Grundlage für die Untersuchung von Suffixen.
Die Stichwortgrundlage der großen rückläufigen Wörterbücher (Mater 1967, Muthmann 1991) entspricht in ihrer Größenordnung etwa der Stichwortanzahl des Dudens. Wie im vorigen Abschnitt gezeigt, enthält ein sehr großes Korpus ein Vielfaches der Wortformen eines Wörterbuchs. Am Beispiel der Selbstkomposita konnte der Nutzen dieser großen Wortformenmenge für die Entdeckung von Wörterbuchlücken aufgezeigt werden. Begrenzt wird der Nutzen der Korpora nur durch die Tatsache, dass sich unter den für das Wörterbuch unbekannten Wortformen viele transparente Wortformen befinden, die sich daher nicht für die Aufnahme in ein Wörterbuch eignen, welches nur lexikalisierte Formen aufnimmt. Diese Begrenzung gilt bei einem rückläufigen Wörterbuch nicht mehr unbedingt, zumindest, wenn man es als Reimwörterbuch oder Morphologiegrundlage nutzen möchte. Hier kommt es dem Benutzer auf möglichst große Vielfalt an. Somit müssten sich Korpora als Grundlage für rückläufige Wörterbücher sehr gut eignen. An einem Beispiel[15] haben wir die den Inhalt eines rückläufigen Wörterbuchs mit dem DWDS-Ergänzungskorpus verglichen. Mater (1967) verzeichnet etwa 100 verschiedene Substantive, die auf *–kasten* enden, wohingegen das DWDS-Ergänzungskorpus 1500 Wortformen enthält, die auf *–kasten* enden. Beispiele von Wortformen, die allesamt 25 oder mehr Belegstellen haben, sind:

> *Almosenkasten (28), Baukasten (462), Besteckkasten (25), Betonkasten (43), Bettkasten (91), Bierkasten (52), Blechkasten (67), Blumenkasten (59), Briefkasten (2693), Brotkasten (34), Brustkasten (113), Brutkasten (283), Bühnenkasten (72), Chemiebaukasten (26), Farbkasten (39), Fernsehkasten (40), Flimmerkasten (50), Geigenkasten (88), Glaskasten (515), Guckkasten (303), Haberkasten (29), Handschuhkasten (25), Hausbriefkasten (33), Hirnkasten (28), Holzkasten (161), Instrumentenkasten (80), Karteikasten (100), Kettenkasten (37), Kummerkasten (145), Leierkasten (154), Leuchtkasten (61), Malkasten (121), Metallkasten (40), Modellbaukasten (38), Nistkasten (73), Oberkasten (33), Postkasten (131), Radkasten*

[15] Dieses Beispiel wurde in einem Vortrag von H. M. Enzensberger zum „Tag der Geisteswissenschaften" der BBAW (29.10.2003) angeführt, um die Vorzüge eines rückläufigen Wörterbuchs zu erläutern.

(59), Resonanzkasten (45), Sandkasten (999), Schaltkasten (75), Schaukasten (461), Schwitzkasten (319), Setzkasten (141), Sicherungskasten (79), Sinkkasten (190), Souffleurkasten (69), Spülkasten (49), Starenkasten (59), Stromkasten (43), Stromverteilerkasten (35), Textkasten (1011), Tuschkasten (25), Unterkasten (86), Verbandskasten (64), Verdeckkasten (32), Verteilerkasten (87), Wagenkasten (83), Wasserkasten (30), Werkzeugkasten (244), Wirbelkasten (95), Zahlenkasten (36), Zauberkasten (85), Zettelkasten (331)

Alleine auf *–baukasten* enden im DWDS-Ergänzungskorpus 177 Substantive, also bereits mehr als in Mater unter dem Suffix *–kasten* enthalten sind. 42 Wortformen haben drei oder mehr Belegstellen:

Anker-Steinbaukasten (6), Ankersteinbaukasten (3), Baukasten (462), Bildbaukasten (3), Chemiebaukasten (26), Elektrobaukasten (4), Elektronik-Baukasten (3), Feld1 (Feld2), Genbaukasten (6), Geschichtsbaukasten (3), Grundbaukasten (7), Holzbaukasten (5), Kinderbaukasten (4), Konstruktionsbaukasten (3), Konzern-Baukasten (3), Konzernbaukasten (4), Kosmos-Baukasten (6), Legobaukasten (11), Lego-Baukasten (12), Märklin-Baukasten (6), Märklinbaukasten (3), Märklin-Metallbaukasten (3), Medienbaukasten (5), Metallbaukasten (18), Modellbaukasten (38), Modulbaukasten (4), Opel-Baukasten (3), PSA-Baukasten (3), Satzbaukasten (7), Schatzbaukasten (5), Setzbaukasten (5), Spielbaukasten (5), Spielzeugbaukasten (6), Stabilbaukasten (18), Steckbaukasten (3), Steinbaukasten (14), Technik-Baukasten (4), Technikbaukasten (3), VW-Baukasten (5), Wiederaufbaukasten (3), Zielbaukasten (4), Ziel-Baukasten (3)

Auch unter den Hapax-Legomena finden sich einige bemerkenswerte Komposita, wie beispielsweise der Begriffsbaukasten, zu dem das Korpus den folgenden Beleg liefert:

*„Stolpe greift zielsicher in den **Begriffsbaukasten“***.
Geis, Matthias, Stolpe - die Krönung einer Kampagne, in: die tageszeitung - 12 ½ Jahre taz auf CD-ROM, Berlin: Contrapress-Media-GmbH 1999 [1992]

Korpora zur Ergänzung von Nominalisierungsverbgefügen (NVG)

Diese Verbindungen (z.B. *Hilfe leisten, Erlaubnis erteilen*) bestehen aus einem verbalen Teil (z.B. *leisten, erteilen*), welches in den Grammatiken als Nominalisierungsverb bezeichnet wird, und einem nominalen Teil (*Hilfe, Erlaubnis*). Dieser nominale Teil, oft auch als prädikatives Nomen (abgeleitet vom frz. nom prédicatif) bezeichnet, ist normalerweise ein Nomen actionis, häufig eine Ableitung aus einem Verb oder Adjektiv. NVG-Konstruktionen stehen im Spannungsfeld von idiomatischen Konstruktionen (*auf den Schlips treten*) und „freien“ syntagmatischen Konstruktionen (*in die Pfütze treten*). Sie werden in Wörterbüchern meist unter dem verbalen Teil aufgeführt, obwohl sich die Gesamtbedeutung der Konstruktion aus dem Nomen ableitet. Wörterbücher sollten daher diese prädikativen Nomen möglichst vollständig erfassen. Wir ziehen hierfür als Beispiel das bereits erwähnte Verb *hegen* heran. Das WDG verzeichnet für *hegen* als Nominalisierungsverb 28 Nomen:

Abneigung, Abscheu, Absicht, Achtung, Argwohn, Bedenken, Befürchtungen, Besorgnis, Bewunderung, Ekel , Erwartung, Freundschaft, Furcht, Gedanken, Gefühle, Gesinnung, Groll, Haß, Hoffnung, Liebe, Meinung, Mißtrauen, Plan, Verdacht, Vermutungen, Wunsch, Zorn, Zuneigung, Zweifel.

In Abschnitt 6, Beispiel 2, haben wir gezeigt, dass es nicht gelingt, alle diese Nomen mit statistischen Mitteln zu ermitteln: denn lediglich 14 bzw. 23 der im WDG aufgeführten

prädikativen Nomen sind statistisch signifikante Kollokationspartner im DWDS-Kernkorpus bzw. im DWDS-Ergänzungskorpus.
Für das Verb *hegen* findet man 6661 Belege im DWDS-Ergänzungskorpus. Wir haben eine zufällige Stichprobe von 20% manuell ausgewertet[16]. Nur zwei Nomen, die das WDG aufführt, waren im Korpus nicht vorhanden: *Ekel* und *Zorn*. Umgekehrt konnten wir jedoch im Korpus 72 verschiedene prädikative Nomen finden. Mit anderen Worten sind mehr als 44 Nomen, die im Korpus auftauchen, nicht im WDG verzeichnet. Die in Klammern angegebenen Zahlen entsprechen der Häufigkeit in der Stichprobe:

Sympathie (25) Vorstellung (17) Illusion (15) Vertrauen (13) Traum (13) Glauben (11) Vorurteil (10) Vorbehalt (10) Idee (10) Ansicht (9) Auffassung (9) Interesse (8) Vorliebe (7) Neigung (6) Empfindung (6) Ambitionen (6) Anschauung (5) Gelüste (5) Feindschaft (5) Ressentiment (5) Verehrung (5) Zuversicht (4) Skepsis (4) Anspruch (3) Erinnerung (3) Ehrfurcht (3) Verachtung (3) Angst (3) Vision (3) Skrupel (3) Antipathie (2) Ahnung (2) Geringschätzung (2) Hochachtung (2) Affekt (2) Respekt (2) Scheu (2) Stolz (2) Tabu (2) Verlangen (2) Widerwillen (2) Zurückhaltung (2) Kultus (2)

9. Schlussbemerkung

Die geistige Leistung, die hinter den großen einsprachigen Wörterbüchern des Deutschen und anderer vergleichbarer Kultursprachen steht, ist bewundernswert – insbesondere wenn man an die technischen Möglichkeiten denkt, die ihren Bearbeitern zur Verfügung gestanden haben. Aber sie sind alles andere als vollkommen. Der gezielte Einsatz großer, ausgewogener Korpora macht es nunmehr möglich, einigen diesen Unvollkommenheiten beizukommen und eine bessere Abdeckung des lexikalischen Repertoires einer Sprache zu erreichen, die ohne diese Möglichkeit illusorisch wäre.

Summary

Electronic corpora are well-established as a basic resource for dictionary updates. Due to the absence of balanced corpora, updates of German dictionaries depend up to now on opportunistic corpora. This work addresses the question to what extent the newly created balanced DWDS-corpus as well as the very large DWDS-Ergänzungskorpus with a size of 1 billion running words change the picture. We provide evidence for the fact that corpus extraction results vary with the difference in quality and quantity of corpora, and we give several examples for the impact that these corpora potentially have on dictionary updates.

[16] Ich möchte hierbei Anne Urbschat für die Auswertung danken.

Literatur:

Bergenholtz, Henning/ Mugdan, Joachim (1990): Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher. In: Hausmann, Franz Josef/ Reichmann, Oskar/ Wiegand, Herbert Ernst/ Zgusta, Ladislav (eds.): Wörterbücher. Ein internationales Handbuch zur Lexikographie, 2. Teilband. Berlin/New York: de Gruyter [Handbücher zur Sprach- und Kommunikationswissenschaft; 5,2], S. 1611-1625.

Church, Ken; Hanks, Patrick (1990). Using Statistics in Lexical Analysis. In U. Zernik, Hrsg., Exploiting On-line Ressources to build a Lexicon, S. 115-164. Erlbaum: Hillsdale.

Curran, James R.; Miles Osborne. *A Very Very Large Corpus Doesn't Always Yield Reliable Estimates.* In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, 2002.

Dumais, Susan; Michele Banko; Eric Brill, Jimmy Lin and Andrew Ng (2002). Web question answering: Is more always better? In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, S. 291-298. Tampere, Finland.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19m S. 61-74.

Fellbaum Christiane, Hg. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts.

Geyken, Alexander, Alexey Sokirko, Ines Rehbein, Christiane Fellbaum (2004), "What is the optimal corpus size for the study of idioms?", DGfS-Jahrestagung, Mainz 25.-27.02.2004

Hausser, Roland (1998): »Häufigkeitsverteilung deutscher Morpheme.« *LDV-Forum*15.1, S.6-26.

Heid, Ulrich , Worsch, Wolfgang, Evert, Wermke, Dougherty, Vincent (2000), Computational linguistic tools for semi-automatic corpus-based updating of dictionaries . In Proceedings of LREC, 2000.

Heid, Ulrich , Bettina Säuberlich, Esther Debus-Gregor, Werner Scholze-Stubenrecht (2004). Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition. In: LREC 2004 Proceedings, Lissabon (Portugal), S. 911-914.

Kilgarriff, Adam, Tugwell, David. Word Sketch: Extraction and Display of Significant Collocations for Lexicography". In Proceedings of the workshop "Collocation, Computational Extraction, Analysis and Exploitation", 39th ACL & 10th EACL, Toulouse, July 2001, A. 32-38.

Klein, W., and A. Geyken (2000), Projekt "Digitales Wörterbuch der deutschen Sprache des 20. Jh.". In: Jahrbuch der BBAW 1999, Berlin: Akademie Verlag, S. 277-289.

Klein, Wolfgang (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In J. Scharnhorst, Hrsg., *Sprachkultur und Lexikographie*. Frankfurt/M: Peter Lang, S. 281 - 309.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. German Research Center for Artificial Intelligence and Saarland University Dissertations. In Computational Linguistics and Language Technology, Volume 7. Saarbrücken.

Kučera, H. und W. N. Francis (1967). Computational analysis of present-day English. Brown University Press. Providence, Rhode Island.

Kunze, Claudia (2000). Extension and use of GermaNet, a lexical-semantic database. In: *Proceedings of the Second International Conference on Language Resources and Evaluation, Athens*, Vol. II, S. 999–1002.

Manning, Christopher, Hinrich Schütze (2000). Foundations of statistical natural language processing. Cambridge, London.

Mater, Erich (1967): Rückläufiges Wörterbuch der deutschen Gegenwartssprache. Leipzig.
Muthmann, Gustav (1991): Rückläufiges deutsches Wörterbuch. Handbuch der Wortausgänge im Deutschen, mit Beachtung der Wort- und Lautstruktur. 2., unveränd. Aufl. Tübingen.
Mikheev A., M. Moens, and C. Grover. (1999). Named Entity recognition without gazetteers. In Proc. of EACL, Bergen, Norway. EACL
Rieger, Burkhard (1979): Repräsentativität: Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In: Bergenholtz, Henning/ Schaeder, Burkhard (eds.): Empirische Textwissenschaft. Aufbau und Auswertung von Textkorpora. Königstein/Ts.: Scriptor [Monographien Linguistik und Kommunikations-wissenschaft; 39], S. 52-70.
Rundell, Michael (1996). 'The corpus of the future, and the future of the corpus'. Special conference on 'New Trends in Reference Science'. Exeter.
Rundell, Michael (2000). "The biggest corpus of all", Humanising Language Teaching. Year 2; Issue 3; May 2000.
Senellart, Jean. (1996), "Statistique Prudence: Quelques études statistiques avec les noms-composés dans le Journal Le Monde". 3e Journées d'Intex. Paris. Jussieu.