3. Analyse von Korpusdaten

3.1 Beschreibungsebenen

Um das sprachliche Material in einem Korpus klassifizieren, analysieren und interpretieren zu können, ist es sinnvoll, sich vorab über die verschiedenen Beschreibungsebenen klar zu werden. Zu unterscheiden sind dabei nicht nur die verschiedenen korpuslinguistischen Kategorien der Textwörter, Tokens und Types, wichtig ist auch die Abgrenzung der korpuslinguistischen Begriffe Textwort, Wortform-Type und Lemma-Type von den linguistischen Beschreibungskategorien Wort, Wortform und Lexem. Betrachten wir dazu erst einmal ein Beispiel (vgl. Linke *et al.* 2004). Wie viele Wörter enthält der folgende Satz:

(5) Wenn hinter Fliegen eine Fliege fliegt, fliegt eine Fliege Fliegen hinterher.

Hier gibt es mehrere richtige Antworten: elf, sieben oder sechs. Versteht man Wörter als Einheiten der Schriftsprache (orthografische Wörter), so muss die Antwort elf lauten, denn der Satz enthält elf Einheiten, die durch Leerzeichen oder Satzzeichen voneinander getrennt sind:

(6) 1. Wenn, 2. hinter, 3. Fliegen, 4. eine, 5. Fliege, 6. fliegt, 7. fliegt, 8. eine, 9. Fliege, 10. Fliegen, 11. hinterher

In einem Korpus werden solche, mithilfe der Schreibung definierten Worteinheiten als **Textwörter**, Wortform-Tokens oder laufende Wortformen bezeichnet. Definiert man Wörter hingegen als syntaktische Wörter oder Wortformen, d.h. als formal voneinander unterscheidbare Bestandteile des Satzes, so kommt man nur noch auf sieben Wörter:

(7) 1. wenn, 2. hinter, 3. Fliegen, 4. eine, 5. Fliege, 6. fliegt, 7. hinterher

Die Formen *Fliegen*, *eine*, *Fliege* und *fliegt* kommen im Satz doppelt vor. Sie werden bei dieser Betrachtung zusammengefasst und zählen jeweils als eine Wortform, da sie sich in ihrer Form nicht unterscheiden. In einem Korpus werden Worteinheiten, die über Unterschiede in ihrer Form definiert werden, als Wortform-Types bezeichnet. Schließlich kann man sich fragen, ob es sinnvoll ist, die Flexionsformen *Fliege* und *Fliegen* als zwei getrennte Wörter zu behandeln. Der einzige Unterschied zwischen den beiden Wortformen besteht darin, dass *Fliegen* eine zusätzliche grammatische Markierung für den Plural trägt, nämlich das Flexionssuffix *-n*, während die Wortform *Fliege* keine zusätzliche Markierung für den Singular hat. In beiden Fällen handelt es sich aber um Vertreter derselben Tierart. Anders wäre das etwa im Fall von (8).

(8) Wenn hinter Fliegen eine Biene fliegt, fliegt eine Biene Fliegen hinterher.

Abstrahiert man also von grammatischen Markierungen, so gehören sowohl die Wortform *Fliege* als auch die Wortform *Fliegen* zum selben Wort, nämlich dem Lexem oder morphologischen Wort *FLIEGE*. Zählt man nun also die Lexeme im Beispielsatz, so kommt man auf sechs Wörter: die Konjunktion *WENN*, die Adverbien *HINTER* und *HINTERHER*, das Nomen *FLIEGE*, den Artikel *EINE* und das Verb *FLIEGEN*. In einem Korpus werden Worteinheiten, die sich nur in ihren Flexionsmerkmalen unterscheiden, unter dem Begriff Lemma oder Lemma-Type zusammengefasst. Um im Folgenden auch optisch zu verdeutlichen, wenn explizit von Lemma-Types bzw. Lexemen die Rede ist, werde ich diese im Text in Kapitälchen angeben (*FLIEGE*).

Die bisher eingeführten Begriffe Lemma-Type, Wortform-Type und Wortform-Token haben jedoch einen Nachteil: sie beziehen sich auf Worteinheiten. Korpusanalysen sind jedoch nicht auf die Untersuchung von Worteinheiten beschränkt, sondern können auch auf Laut-, Satz-, Text- oder Bedeutungsebene durchgeführt werden. Aus diesem Grund unterscheidet man in der Korpuslinguistik üblicherweise unabhängig von der Sprachebene zwischen Types und Tokens. Bei einem **Token** handelt es sich ganz allgemein um das konkrete Vorkommen einer sprachlichen Einheit im Korpus. Das kann eine bestimmte Wortform, Lautäußerung oder Phrase sein. Ein **Type** ist hingegen die abstrakte sprachliche Einheit, die zusammengehörige Tokens wie Wortformen oder Lautvarianten zusammenfasst und dabei von konkreten Merkmalsausprägungen wie Flexions- oder Intonationsmerkmalen abstrahiert.

Verdeutlichen wir uns die Unterscheidung zwischen Types und Tokens an einem konkreten Beispiel, nämlich der Untersuchung von Wortbildungswandel im Mainzer Zeitungskorpus (Scherer 2005). Ziel der Untersuchung war es festzustellen, ob und wie sich die Möglichkeiten zur Bildung von Wörtern in den letzten vier Jahrhunderten verändert haben. Dazu wurden Nomen wie *Lehrer*, *Schüler* und *Lacher* untersucht, die mithilfe des Wortbildungssuffixes *-er* aus Verben (*lehren*, *lachen*), Nomen (*Schule*) und anderen Wortarten abgeleitet wurden. Der folgende Ausschnitt stammt aus dem ersten von insgesamt neun Teilkorpora.

Den 12. diß hat der Herr von Veneſi zu Brůſſel ein Pancket vnd Dantz gehalten / vnd ſeinem **Diener** oder **Gertner** befohlen / niemand ohn ſeinen wiſſen in den Saal zulaſſen / gleich hernach iſt ein Spanniſcher Hauptman vnd **Ritter** *Don Roderico Floris* genant / aber nur ſchlecht / wie ein **Diener** bekleidt für den Saal kommen / vnd ſich mit gewalt eindringen wollen / dem **Gertner** zwey Maultaschen etliche Gemechtſtöß / vnd ein Stich in Arm geben / darauff der **Diener** im 3orn ſeinen Dolchen außgezogen / vnd den Hauptman alßbald erſtochen / welchen man hernach in des Spinnola Hauß geführet / vnnd den **Gertner** einziehen laſſen / vnd die Freud eingestellt worden / als nun den dritten Tag / der Hauptman von allen *Officirn* vnd Herrn von Hoff ins Cloſter zu den **Auguſtinern** zur begrebnis begleitet / iſt entzwiſchen der **Theter** zum Galgen geführt / gehenckt vnd ihme die rechte Hand abgehawen / vnd an Galgen genagelt worden / das gedůnckt jederman ein frembder Sententz ſein / weil der **Theter** ſeinem Befehlig nachkommen / vnd ſich auch der Nothwehr gebrauchen můſſen.

Abbildung 3: Mainzer Zeitungskorpus: Teilkorpus 1609 *Aviso* (Ausschnitt)

Analysieren wir den Ausschnitt zunächst im Hinblick auf die Bildung von Personenbezeichnungen mit dem Suffix *-er*. Der Ausschnitt enthält insgesamt zehn Tokens (vgl. 9a), die im Text fett markiert sind. Diese zehn Tokens verteilen sich auf insgesamt fünf Types (vgl. 9b).

(9) a. Diener, Gertner, Ritter, Diener, Gertner, Diener, Gertner, Auguſtinern, Theter, Theter
b. AUGUSTINER (1 Token), DIENER (3 Tokens), GÄRTNER (3 Tokens), RITTER (1 Token), TÄTER (2 Tokens)

Untersuchen wir den Ausschnitt hingegen im Hinblick auf eine andere Fragestellung, etwa die Verwendung von Präpositionen, so kommen wir zu einem anderen Ergebnis. Der Ausschnitt enthält in diesem Fall 16 Tokens (vgl. 10a), die im Text unterstrichen sind. Sie verteilen sich auf sieben Types (vgl. 10b).

(10) a. von, zu, ohn, in, für, mit, in, im, in, von, von, ins, zu, zur, zum, an
b. AN (1 Token), IN (5 Tokens), MIT (1 Token), OHNE (1 Token), VON (3 Tokens), VOR (= für) (1 Token), ZU (4 Tokens)

Aufgabe 10: Wie viele Textwörter enthält der oben stehende Ausschnitt aus dem Mainzer Zeitungskorpus? Bitte ermitteln Sie die Zahl der Types (Lemma-Types) und Tokens für die im Ausschnitt enthaltenen Nomen. Zählen Sie Eigennamen zu den Nomen.

Wie gesagt befasst sich aber nicht jede korpuslinguistische Analyse mit einer Untersuchung auf Wortebene. Vielmehr gibt es Fragestellungen, die sprachliche Einheiten betreffen, die größer oder kleiner sind als das Wort. So erfolgt die bereits erwähnte Untersuchung von Elter (2005) zur Kasusverwendung bei *wegen* auf der

syntaktischen Ebene der Phrase. Hier werden die Begriffe Type und Token auf die im Korpus enthaltenen *wegen*-Phrasen angewendet. Um Tokens handelt es sich also etwa bei den Phrasen *wegen des Mondscheinfrisierens* oder *wegen dem starken Wind* (vgl. Kapitel 1.2). Diese Tokens verteilen sich in Elters Studie auf zwei zugrunde liegende grammatische Muster, *wegen* + Genitiv und *wegen* + Dativ. Diese beiden Muster stellen die Types dar.

Untersucht man hingegen wie Dittmar/Bressem (2005) die Verbstellung in Nebensätzen mit *weil*, so bezieht sich die Zahl der Tokens auf die Anzahl der *weil*-Nebensätze und die der Types auf die beiden Verbstellungsvarianten: finites Verb an letzter Stelle wie in (11a) bzw. finites Verb an zweiter Stelle wie in (11b).

(11) a. weil ich das immer so **mache**
b. weil das **mache** ich immer so

Festzuhalten sind demnach zwei Dinge: Zum einen kann sich die Grundgesamtheit, auf die sich die Begriffe Type und Token beziehen, je nach Fragestellung verändern. Dahingegen ist die Anzahl der Textwörter in einem Korpus unabhängig von der untersuchten Fragestellung. Zum anderen haben die Begriffe Type und Token nicht zwangsläufig einen Bezug zur Wortebene. Vielmehr können die Begriffe auf sprachliche Einheiten unterschiedlicher Ebenen wie Wort, Satz, Text bzw. deren Bestandteile verweisen.

3.2 Methoden

Korpora können sowohl qualitativ als auch quantitativ ausgewertet werden. Der wesentliche Unterschied zwischen qualitativen und quantitativen Korpusanalysen besteht nicht darin, welche Fragestellungen untersucht werden, sondern wie diese untersucht werden. Nehmen wir das Beispiel Fremdwörter. In zwei unterschiedlichen Studien analysieren Schanke (2001) und O'Halloran (2002) den Einfluss englischer und französischer Entlehnungen, so genannter Anglizismen und Gallizismen, im Deutschen. Beide arbeiten mit einem selbst zusammengestellten Korpus aus Zeitungen bzw. Zeitschriften. Schankes Korpus enthält sämtliche Ausgaben des *Handelsblatts* aus dem März 2000, O'Hallorans Korpus umfasst ein Teilkorpus zur Modesprache mit mehreren Jahrgängen der Frauenzeitschrift *Brigitte* sowie ein Teilkorpus zur Standardsprache, das mehrere Jahrgänge des Nachrichtenmagazins *Stern* und der *Berliner Illustrierten Zeitung* enthält.

Abgesehen von der Größe der Korpora unterscheiden sich die beiden Studien auch in ihrer Methode. Während Schanke in seiner Korpusanalyse einen qualitativen Ansatz wählt, untersucht O'Halloran ihr Korpus unter quantitativen Gesichtspunkten.

Schankes Ziel ist es, die gefundenen Fremdwörter im Hinblick auf ihre Wortart zu klassifizieren und sie bestimmten Themenbereichen wie Computerbranche, Börse oder Bankwesen zuzuordnen. Bei Schankes Untersuchung geht es also darum, in einem Korpus die Existenz bestimmter sprachlicher Erscheinungen, nämlich Anglizismen, festzustellen, die einzelnen Wörter herauszusuchen und sie nach bestimmten Kriterien, konkret nach Wortfeldern, zu klassifizieren. Schankes Vorgehen entspricht dem einer **qualitativen Korpusanalyse**. Qualitative Korpusanalysen legen ihren Schwerpunkt auf die Ermittlung, die Klassifizierung, die Einordnung und Interpretation von bestimmten Phänomenen.

Im Gegensatz dazu steht O'Hallorans Arbeit. O'Halloran untersucht die Verbreitung von englischen und französischen Fremdwörtern innerhalb der letzten einhundert Jahre. Dabei stellt sie fest, dass der Anteil an Fremdwort-Types im Gesamtkorpus steigt, und zwar von 0,6% im Jahr 1902 auf 2,0% im Jahr 1997. Darüber hinaus beobachtet sie, dass der Anteil von Fremdwort-Tokens im Teilkorpus zur Modesprache zu jedem Zeitpunkt den Fremdwortanteil im Teilkorpus zur Standardsprache übersteigt. Im Jahr 1997 liegt der Fremdwortanteil in der Standardsprache z.B. bei 4,0%, in der Modesprache hingegen bei 14%. O'Halloran geht es in ihrer Untersuchung also darum, die **Frequenz** von bestimmten Phänomenen zu ermitteln und miteinander zu vergleichen, um daraus Rückschlüsse über die untersuchte Fragestellung ziehen zu können. Das Bestimmen von Häufigkeiten im Korpus und die sich daraus ergebende Möglichkeit, Ergebnisse unmittelbar miteinander zu vergleichen, ist das Kennzeichen **quantitativer Korpusuntersuchungen**.

An quantitativen Kennzahlen wird standardmäßig die Korpusgröße ermittelt, die üblicherweise in Textwörtern gemessen wird (vgl. Kapitel 4.8). Sie bildet die wichtigste Bezugsgröße für alle quantitativen Auswertungen. Ist die Größe eines Korpus nicht bekannt, sind quantitative Analysen nur dann sinnvoll, wenn die Ergebnisse für mehrere ähnlich geartete Phänomene innerhalb des Korpus verglichen werden können.

Von besonderem Interesse ist für den Forscher die Anzahl der Types und Tokens des untersuchten Phänomens, da diese beiden Zahlen Auskunft darüber geben, wie oft ein Phänomen insgesamt im Korpus belegt ist (Tokens) und auf wie viele unterschiedliche

Ausprägungen des Phänomens (Types) sich die Tokens verteilen. So fanden sich im obigen Ausschnitt aus dem Mainzer Zeitungskorpus insgesamt fünf Types und zehn Tokens für die untersuchten nominalen *-er*-Derivate (vgl. Abbildung 3).

Wichtig ist, die Zahl der Types und Tokens jeweils im Verhältnis zur Korpusgröße zu sehen. Zehn Tokens in einem kleinen Korpus können relativ gesehen eine höhere Frequenz darstellen als hundert Tokens in einem großen Korpus. Liegt die Zahl der Types und Tokens vor, kann man daraus das Verhältnis von Types zu Tokens berechnen. Dieses **Type-Token-Verhältnis** gibt Auskunft darüber, wie viele Tokens durchschnittlich auf einen Type entfallen. Liegt die Anzahl der Tokens je Type sehr hoch, handelt es sich bei den meisten Tokens vermutlich um häufig verwendete Ausdrücke, die eine gewisse Formelhaftigkeit aufweisen. Die Anzahl der spontanen, neuen Formen, die dem untersuchten Muster folgen, ist dann gering. Umgekehrt ist ein niedriges Verhältnis von Types zu Token ein Indiz dafür, dass viele Types nur selten vorkommen. Kommt ein Type nur ein einziges Mal im Korpus vor, spricht man von einem **Hapax Legomenon**. Enthält ein Korpus viele Hapax Legomena und andere seltene Types, ist die Wahrscheinlichkeit hoch, dass das untersuchte sprachliche Muster von den Sprechern bzw. Schreibern produktiv eingesetzt wird und dass nach seinem Vorbild neue Bildungen vorgenommen werden. Allgemein kann der Anteil der Hapax Legomena an der Zahl der Tokens dazu verwendet werden, die **Produktivität** eines sprachlichen Musters zu bestimmen. Je höher der Anteil der Einmalbelege, desto höher ist die Wahrscheinlichkeit, dass das Muster Neubildungen hervorbringt.

Zielt eine Korpusuntersuchung weniger auf die Wortebene als auf die Satzebene ab, so ist es sinnvoll, die Anzahl der Sätze in einem Korpus oder in einem Text zu ermitteln. Zudem können die durchschnittliche Satzlänge, die Zahl der Sätze mit einer bestimmten Länge sowie der Anteil von Sätzen mit einer bestimmten Zahl an Wörtern dazu dienen, syntaktische Charakteristika eines Korpus, eines Textes oder einer Varietät herauszuarbeiten.

Um zu gewährleisten, dass es sich bei den ermittelten Ergebnissen nicht um bloßen Zufall handelt, empfiehlt es sich, die Ergebnisse statistisch abzusichern. Dies geschieht mittels eines Signifikanztests, der sicherstellt, dass die Ergebnisse nicht allein dem Zufall geschuldet sind. Die meisten Signifikanztests gehören jedoch der höheren Mathematik an, sodass es sich empfiehlt, entsprechende Statistikprogramme zu benutzen.

3.3 Vergleichbarkeit von Daten

Beim Vergleich von Daten aus unterschiedlichen Korpora ist es wichtig, qualitative und quantitative Charakteristika der Korpora zu beachten. Zum einen sollte überlegt werden, ob sich die Korpora aufgrund ihrer Konzeption überhaupt vergleichen lassen und wenn ja, in welchem Rahmen. Auf den ersten Blick scheint es wenig sinnvoll zu sein, ein Korpus der Kindersprache mit einem historischen Korpus oder einem Korpus zur Fachsprache der Biologie zu vergleichen. Dennoch kann ein solcher Vergleich sinnvoll sein, wenn untersucht werden soll, ob Parallelen in der kindlichen und der historischen Sprachentwicklung bestehen oder ob Biologiebücher für die Schule den Entwicklungsstand der Kinder angemessen berücksichtigen, was die Bezeichnung von Pflanzen, Tieren und deren Teilen betrifft.

Als Vergleichsgrundlage dienen häufig Referenzkorpora, die als Standard verwendet werden, um Abweichungen zwischen Varietäten und Standardsprache festzustellen (vgl. Kapitel 2.9). Ein Referenzkorpus kann also dazu dienen, festzustellen, inwieweit sich das Bairische, die Fachsprache der Medizin oder das Mittelhochdeutsche von der Standardsprache unterscheiden. Daneben ist es aber auch wichtig, auf die quantitative Vergleichbarkeit zu achten. Tabelle 1 zeigt die Ergebnisse einer Suchabfrage in drei verschiedenen Korpora des IDS, dem Bonner Zeitungskorpus, dem LIMAS-Korpus und dem Mannheimer Korpus 1. Gesucht wurde nach den Wortformen *Buch*, *Hochhaus* und *Universität*.

Suchbegriff	Bonner Zeitungskorpus	LIMAS-Korpus	Mannheimer Korpus 1
Buch	**313**	166	229
Hochhaus	**16**	3	6
Universität	**315**	116	219

Tabelle 1: Ergebnisse der Stichwortsuche (absolut)

Wie man sieht, finden sich im Bonner Zeitungskorpus jeweils die meisten und im LIMAS-Korpus jeweils die wenigsten Belege für alle drei Suchbegriffe. Woran liegt das? Nun, zum einen könnte es an der Zusammensetzung der Korpora liegen: Während das Bonner Zeitungskorpus ausschließlich Zeitungstexte enthält, ist der Anteil an Zeitungstexten in den anderen beiden Korpora gering. Sie bestehen überwiegend aus Textsorten wie Belletristik, Gebrauchsliteratur und wissenschaftlichen Texten. Eine mögliche Folgerung ist also, dass sich alle drei Suchbegriffe überdurchschnittlich häufig in Zeitungstexten finden. Bevor man jedoch einen solchen Schluss zieht, sollte man einen Blick auf die Größe der Korpora werfen, die verglichen werden sollen.

Aufgabe 11: In einem Korpus A finden sich 80 Belege für das Wort *BLUMENTOPF*, in Korpus B 100 Belege für dasselbe Wort. Korpus A und Korpus B enthalten je eine Million Textwörter. Korpus C enthält ebenfalls 100 Belege für *BLUMENTOPF*, aber anderthalb Millionen Textwörter. In welchem der drei Korpora finden sich die meisten Belege für das Wort *BLUMENTOPF*?

Obwohl sich im Bonner Zeitungskorpus mehr Belege für die Wortform *Buch* finden als in den anderen beiden Korpora, bedeutet dies nicht unbedingt, dass *Buch* im Bonner Zeitungskorpus häufiger ist als im LIMAS-Korpus oder dem Mannheimer Korpus 1. Dies liegt daran, dass eine Aussage über die Häufigkeit eines Wortes immer im Verhältnis zur Größe des Korpus gesehen werden muss.

Ein direkter Vergleich von Korpusdaten ist aufgrund unterschiedlicher Korpusgröße im Normalfall nicht möglich. Insofern ist es beim Vergleich von Daten aus verschiedenen Korpora von größter Wichtigkeit, die jeweiligen Ergebnisse ins Verhältnis zur Korpusgröße zu setzen. Geht es darum, die Frequenz bestimmter Wörter anzugeben, so kann dies wie bei O'Halloran (2002) in Form von Prozentangaben geschehen, die sich auf die Zahl der Textwörter oder Lemma-Types im Korpus beziehen. Befasst sich die Untersuchung hingegen nicht mit Einheiten der Wortebene, so kommt nur eine **Normalisierung** infrage. Bei der Normalisierung werden die Ergebnisse auf eine bestimmte Anzahl von Textwörtern, etwa 10.000 oder eine Million, umgerechnet. Dabei sollte sich die Normalisierung an der typischen Textlänge im Korpus orientieren.

Vergleichen wir das Bonner Zeitungskorpus mit dem LIMAS-Korpus und dem Mannheimer Korpus 1, so ergibt sich folgendes Bild: Das Bonner Zeitungskorpus enthält über 3,6 Millionen Textwörter und ist damit fast dreimal so groß wie das LIMAS-Korpus mit rund 1,2 Millionen Textwörtern und etwa anderthalbmal so groß wie das Mannheimer Korpus 1, das rund 2,6 Millionen Textwörter beinhaltet. Es ist also nicht verwunderlich, dass sich im größten Korpus die meisten Belege finden und im kleinsten Korpus die wenigsten! Als Konsequenz aus diesen Größenunterschieden müssen sämtliche Ergebnisse auf eine genormte Korpusgröße umgerechnet werden (vgl. Tabelle 2). Sinnvoll erscheint in diesem Fall eine Normgröße von einer Million Textwörtern.

Suchbegriff	Bonner Zeitungskorpus	LIMAS-Korpus	Mannheimer Korpus 1
Buch	86	**135**	89
Hochhaus	**4**	2	2
Universität	87	**94**	85

Tabelle 2: Ergebnisse der Stichwortsuche (normalisiert je 1 Mio. Textwörter)

Die normalisierten Daten in Tabelle 2 zeigen im Vergleich zu Tabelle 1 ein ganz anderes Bild: Lediglich die Wortform *Hochhaus* kommt mit vier Belegen je Million Textwörter im Bonner Korpus häufiger vor als in den anderen beiden Korpora. Dahingegen finden sich die meisten Belege für *Buch* (135) und *Universität* (94) im LIMAS-Korpus. In den anderen beiden Korpora haben *Buch* und *Universität* hingegen fast dieselbe Frequenz.

Nach dem Vergleich der normalisierten Ergebnisse lautet die Frage also nicht mehr, warum die Wortformen *Buch*, *Hochhaus* und *Universität* im Bonner Zeitungskorpus am häufigsten sind, sondern vielmehr, warum *Buch* im LIMAS-Korpus deutlich öfter vorkommt als in den anderen Korpora, warum *Hochhaus* im Bonner Zeitungskorpus doppelt so oft belegt ist wie in den anderen Korpora und warum *Universität* in allen drei Korpora fast gleich häufig vorkommt.

Bei dem Vergleich von Daten aus verschiedenen Korpora sind demnach zwei Dinge wichtig: Zum einen sollte man sich fragen, ob es im Hinblick auf die zu untersuchende Fragestellung überhaupt sinnvoll ist, zwei gegebene Korpora miteinander zu vergleichen. Zum anderen ist es unerlässlich, bei einem Vergleich von korpusbasierten Häufigkeiten die Korpusgröße zu berücksichtigen, da bei unterschiedlich großen Korpora andernfalls die Ergebnisse des Vergleichs verfälscht werden.

Aufgabe 12: Unten finden Sie die Ergebnisse aus dem Erlanger Dürer-Korpus (Müller 1993) und dem Würzburger Korpus der Wissensliteratur (Brendel *et al.* 1997) zur Wortbildung in frühneuhochdeutschen Fachtexten.
In welchem der beiden Korpora finden sich die meisten Nominalisierungen mit den Suffixen *-er*, *-heit/-keit* und *-ung* (Types, Tokens)? Bitte berechnen Sie zudem das Type-Token-Verhältnis für die einzelnen Suffixe und vergleichen Sie die Ergebnisse miteinander.

Korpus	Text-wörter	*-er*		*-heit/-keit*		*-ung*	
		Types	Tokens	Types	Tokens	Types	Tokens
Dürer-Korpus	440.000	93	700	76	326	193	2.443
Wissens-literatur	1.073.000	510	4.505	454	6.575	1.025	5.213

3.4 Stichwortsuche – die Suche nach Wörtern, Wortformen und Wortteilen

Die einfachste Möglichkeit an Informationen in einem Korpus zu kommen, ist die Suche nach einem bestimmten Wort, einer Wortform oder einem Wortteil wie *HAUS*, *liest* oder *un-*. Genau diese Möglichkeit der Stichwortsuche haben Günther (2002) und Hämmer (2001) genutzt, um die Verwendung des Wortes *stolz* bzw. des Wortteils *-park* zu analysieren.

Anlass für Günthers Untersuchung des Wortes *stolz* war die öffentliche Diskussion um den Satz *Ich bin stolz darauf, ein Deutscher zu sein*, den ein Politiker in einem Interview geäußert hatte. Günther wollte jenseits der gesellschaftlichen Debatten klären, in welchem Zusammenhang das Wort *stolz* verwendet wird. Stolz sein, so ergab Günthers Recherche in den Textkorpora des IDS, kann man nicht nur auf eine Leistung (vgl. 12a), eine berufliche oder private Tätigkeit (vgl. 12b), sondern auch auf eine bestimmte nationale oder geografische Herkunft (vgl. 12c).

(12) a. stolz darauf, Abgeordneter geworden zu sein
b. stolz darauf, ein Bauer/ein Zeitungsleser zu sein
c. stolz darauf, ein Schweizer/ein Münchner zu sein

Wie Günther feststellt, bringt die Äußerung *stolz darauf, ein X zu sein* somit zwar ein gewisses Maß an Selbstbewusstsein zum Ausdruck, sie muss jedoch nicht zwangsläufig ein Zeichen von Überheblichkeit seitens des Sprechers sein.

Anders als Günther suchte Hämmer nicht nach vollständigen Wörtern, sondern lediglich nach einem Wortbestandteil. Gegenstand ihrer Analyse bildeten Komposita mit dem Zweitglied *-park*. Hämmers Suche im Korpus des Projekts Deutscher Wortschatz in Leipzig ergab, dass die Komposita mit *-park* semantisch in zwei Gruppen zerfallen. Bei der größeren Gruppe handelt es sich um klassische Determinativkomposita, bei denen *-park* als Grundwort auftritt (vgl. 13a).

(13) a. Schlosspark, Tierpark, Vergnügungspark
b. Gerätepark, Unternehmenspark, Windpark

Ein *Schlosspark*, *Tierpark* oder *Vergnügungspark* ist eine bestimmte Art von Park, die durch das Erstglied näher bestimmt wird: ein Park am Schloss, ein Park mit Tieren, ein Park, den man zur Vergnügung besucht. Daneben fand Hämmer aber auch Beispiele wie in (13b), wo *-park* im Sinne von 'Ansammlung, Gesamtheit von X' interpretiert werden muss. Ein *Gerätepark* ist nicht ein Park

mit Geräten, sondern vielmehr eine Ansammlung von Geräten, ein *Unternehmenspark* nicht der Park eines Unternehmens, sondern eine Ansammlung verschiedener Unternehmen an einem Ort usw. Insofern konnte Hämmer anhand ihrer Korpusanalyse nachweisen, dass sich bei einer Gruppe der *-park*-Komposita die Bedeutung des Zweitglieds semantisch verschiebt, eine Entwicklung, die charakteristisch ist für die Grammatikalisierung von Kompositionsgliedern zu Suffixen.

Die Suche nach einem bestimmten Stichwort ist immer dann sinnvoll, wenn es wichtig ist, unabhängige Informationen über einzelne Merkmale des Suchbegriffs, etwa dessen Bedeutung und Verwendung, zu erhalten. Aus diesem Grund bietet sich die Stichwortsuche in einem Korpus auch für Deutschlerner und Übersetzer an, da das Korpus authentische Verwendungskontexte für den Suchbegriff aufzeigt. Die Wortsuche leistet aber auch bei der Erstellung von Wörterbüchern unverzichtbare Dienste. So stehen die Herausgeber von Neologismenwörterbüchern, d.h. von Wörterbüchern mit "neuen" Wörtern, vor dem Problem festzustellen, welche Wörter in einer Sprache auch tatsächlich neu sind. Tellenbach (2001) berichtet, dass ein Projektteam durch Lektüre rund 5.000 potenzielle Neologismen aufspürte. Jedes dieser Wörter wurde anschließend in den Korpora des IDS überprüft. Die Überprüfung führte zu dem Ergebnis, dass weniger als ein Fünftel der Wörter, die das Wörterbuchteam als Neologismen eingeschätzt hatte, auch tatsächlich neu in der Sprache waren.

Da die Stichwortsuche in einem Korpus in der Regel keinerlei Vorwissen, sondern höchstens etwas Probierfreude erfordert, ist sie weit verbreitet und bietet sich als Einstieg in die Korpusarbeit an. Man gibt die zu suchende Form in die Suchmaschine ein, wartet ab, was passiert, und versucht es gegebenenfalls mit einem leicht abgeänderten Suchwort nochmals. Eine Sache, die man bedenken sollte, ist, dass manche Suchabfragen zwischen Groß- und Kleinschreibung unterscheiden. Teilweise muss man auch bei der Suche nach Wortbestandteilen zusätzliche Platzhalter eingeben.

Prinzipiell ist die Stichwortsuche in Papierkorpora ebenso möglich wie in Computerkorpora. Der grundlegende Unterschied besteht darin, dass bei Papierkorpora die Suche nicht automatisch durchgeführt werden kann, sondern manuell ausgeführt werden muss (vgl. Kapitel 4.6).

Allerdings sind die Möglichkeiten, die die Stichwortsuche bietet, begrenzt. Insbesondere die Homographie und Polysemie von Textwörtern stellen ein Problem dar. Verdeutlichen wir uns das am Beispiel der Homographie (zur Polysemie vgl. Kapitel 3.5). Im Fall von Homographie gehören gleich geschriebene Wortformen wie *Regen* in (14) nicht zum selben Lemma-Type.

(14) a. Trotz strömendem **Regen** blieben die Zuschauer bis zum Ende der spannenden Begegnung.
b. **Regen** Absatz fand auch der neue Kleinwagen des japanischen Autoherstellers.

Während *Regen* in (14a) eine Wortform des Nomens *REGEN* darstellt, handelt es sich bei der identischen Wortform *Regen* in (14b) um eine Form des Adjektivs *REGE*.

Da in einem reinen Textkorpus keine Informationen zur Wortart oder zum zugehörigen Lexem verfügbar sind, operiert eine Suchabfrage in einem solchen Korpus prinzipiell auf der Ebene der syntaktischen Wörter, also der Wortform-Types. Homographe Wortformen wie *Regen* in (14a) und (14b) können in einem nicht annotierten Korpus nicht auseinander gehalten werden. Grammatisch annotierte Korpora hingegen erlauben, bei der Suche zwischen den Lemma-Types *REGEN* und *REGE* zu unterscheiden (vgl. Kapitel 5.3).

Homographe Wortformen und polyseme Wörter können jedoch auch in reinen Textkorpora unterschieden werden, wenn man die einzelne Wortform nicht isoliert, sondern in ihrem Kontext, d.h. in ihrer unmittelbaren Textumgebung, betrachtet. In diesem Fall kann der Kontext die Information ersetzen, die eine Annotation bietet.

Aufgabe 13: Welche Bedeutungen haben die Wörter *Karte*, *decken* und *grün*? Wie werden sie verwendet? Im Zusammenhang mit welchen anderen Wörtern werden sie häufig verwendet? Gibt es feste Redewendungen? Welches der Wörter wird am häufigsten verwendet? Überprüfen Sie Ihre Intuition anhand eines Wörterbuchs und mithilfe einer Suchmaschine im Internet.

3.5 Konkordanzen – Wörter und Wortformen im Kontext

Eine **Konkordanz** ist eine Liste, die alle Vorkommen eines ausgewählten Wortes – oder auch mehrerer Wörter – im Kontext zeigt. Das Wort, für das die Konkordanz erstellt wird, wird auch Knoten genannt. Für Konkordanzen üblich ist eine zeilenweise Darstellung, die als **KWIC** von englisch *key word in context* bezeichnet wird. Dabei wird der Suchbegriff in der Mitte einer Textzeile dargestellt und üblicherweise grafisch hervorgehoben. Auf dessen linker und rechter Seite wird so viel Kontext angegeben, wie die Zeile erlaubt. Abbildung 4 zeigt einen Ausschnitt aus der Konkordanz für *fahren*

im KWIC-Format. Die Konkordanz wurde mit dem Web-basierten Konkordanzprogramm WebConc erstellt (vgl. Kapitel 5.1).

Mit dem Drahtesel durch Köln Fahrrad **fahren** in Köln liegt voll im Trend. Das Fahrrad hat
m Sonntag mit dem Bus 721 zum Flughafen ge**fahren**. Die Haltestelle Bankstraße ist direkt bei
klauen einen vollbeladenen Möbelwagen und **fahren** davon. Aus einer kleinen Probefahrt wird
in ich gern mal mit BVG zum Einkaufen ge**fahren**. Jetzt nehme ich nur noch Auto - ist auf ein
stens einem freien Feld jetzt prima schwarz **fahren**. Oder habe ich was falsch verstanden? Trac
rüher als viele Ihrer Altersgenossen Auto **fahren** zu dürfen. Gehen Sie verantwortungsvoll d
d Ihren Ausweis immer mit, wenn Sie Auto **fahren**. Halten Sie sich unbedingt an die Auflagen
ungen Service Kontakt email Impressum. Wir **fahren** Sie ... ...und Ihr Gepäck... ...hin und zurü
übermüdet sind Gurten Sie sich immer an **Fahren** Sie defensiv und vorausschauend Denken Si
emals von der Straße weg und ins Gelände **fahren**. Aber man fühlt sich in so einem Ding ein

Abbildung 4: WebConc: Konkordanz für *fahren* (Ausschnitt)

Neben Konkordanzen im KWIC-Format werden auch Konkordanzen verwendet, die ganze Sätze oder Abschnitte des Kontextes oder eine vorher bestimmte Anzahl an Textzeilen wiedergeben.

Konkordanzen erlauben es, die Suchbegriffe in ihrem Kontext zu analysieren. Der Ausschnitt aus der Konkordanz für *fahren* in Abbildung 4 lässt verschiedene Verwendungen des Verbs erkennen, die ohne Kontext nicht zu unterscheiden wären. So verzeichnet die Konkordanz folgende Verwendungen von *fahren*, die im Satz jeweils unterschiedliche syntaktische Ergänzungen bzw. Angaben erfordern:

(15) a. *fahren* mit einem Adverbial der Art und Weise (*Fahren Sie defensiv*)
b. *fahren* mit Akkusativ-Objekt ('etwas fahren', z.B. ein Auto oder ein Fahrrad)
c. *fahren* mit Dativ-Objekt ('jemanden fahren')
d. *fahren* mit einem Adverbial der Richtung ('irgendwohin fahren', z.B. zum Flughafen oder ins Gelände)

Zudem findet sich ein Beleg für die idiomatische Wendung *schwarz fahren*, die nicht wörtlich, sondern im übertragenen Sinn zu verstehen ist als 'fahren ohne Fahrkarte'. Konkordanzen ermöglichen es also, verschiedene Bedeutungen eines Wortes zu erkennen oder bestimmte grammatische Strukturen zu ermitteln, in denen ein Wort verwendet werden kann.

Kehren wir an dieser Stelle zurück zum Problem von Homographie und Polysemie. Mit dem Problem der Polysemie, d.h. der Mehrdeutigkeit von sprachlichen Ausdrücken, hat sich Haß-Zumkehr (2002) beschäftigt (zur Homographie vgl. Kapitel 3.4). Sie interessierte sich für die verschiedenen Bedeutungsvarianten von *Absatz*, wie sie sich in Wörterbüchern und Textkorpora finden. Wörterbücher nennen unter dem Stichwort *Absatz* üblicherweise folgende Bedeutungen:

(16) a. Teil eines Textes, insbesondere eines Gesetzestextes
b. Unterbrechung einer Fläche, etwa einer Treppe oder Mauer
c. Teil eines Schuhs
d. Ablagerung, etwa von Kalk oder Kies
e. Verkauf von Waren und Produkten

Für den korpusbasierten Teil ihrer Untersuchung verwendete Haß-Zumkehr eine Konkordanz, die auf den Textkorpora des IDS basiert. Diese ist ausschnittsweise in Abbildung 5 wiedergegeben:

wie Senatoren hat – bestimmt Artikel 51 **Absatz** 2 des Grundgesetzes : .
die Schuhfabrik hat keinen **Absatz** mehr , beim Glühlampenhersteller
Verteidigungswaffen fänden " reißenden **Absatz** " , berichtete das Hallesche Boule
fkleber fanden unter Trabifahrern guten **Absatz** .
ltog sich im Schatten des Artikels 20, **Absatz** 2 der Verfassung , der die Diskrim
r Forschung über die Produktion bis zum **Absatz** - weltmarktfähige Erzeugnisse her
ndung von Wissenschaft , Produktion und **Absatz** in diesen starken ökonomischen Ei

Abbildung 5: Konkordanz für *Absatz* (Haß-Zumkehr 2002)

In diesem Ausschnitt aus der Konkordanz finden sich Beispiele für die zwei häufigsten Bedeutungsvarianten, die Haß-Zumkehr in ihrem Korpus fand. Dies sind zum einen die wirtschaftsbezogene Lesart von *Absatz* im Sinne von (16e), die die deutliche Mehrheit aller Vorkommen ausmachte. Zum anderen ist es die textbezogene Lesart in (16a), die im Korpus ebenfalls relativ häufig vorkam. Eher selten belegt war hingegen eine dritte Bedeutung von *Absatz*, nämlich 'Teil eines Schuhs' (vgl. 16c). Zwei weitere im Wörterbuch angegebene Bedeutungen, *Absatz* im Sinne von (16b) und (16d), waren in den Konkordanzen nicht zu ermitteln. Wie der Fall von *Absatz* zeigt, kann die Analyse von Konkordanzen somit dazu beitragen, Wörterbucheinträge benutzerfreundlicher zu gestalten.

Konkordanzen machen es möglich, ein bestimmtes Wort in einer Vielzahl von Kontexten zu untersuchen und so eventuelle Bedeutungsvarianten zu erfassen. Wie wir am Beispiel von *fahren* gesehen haben, geben Konkordanzen aber auch Auskunft über den Bedeutungszusammenhang, in dem ein Wort verwendet wird, und über dessen grammatische Einbindung im Satz. Ähnliche Gründe sind auch ausschlaggebend dafür, mehrsprachige Konkordanzen bei der Übersetzung zu nutzen. Wie die deutsch-englische Konkordanz in Abbildung 6 zeigt, lautet die englische Entsprechung für *Sprache* je nach Kontext einmal *power of speech* und einmal *language*.

ac/decker – 632
I could hardly raise my hands; I had lost the power of speech.
Ich konnte kaum noch die Hände heben; ich hatte die **Sprache** verloren.

ac/harm – 478
Some were for revolution, others for reform, most preferring to speak in revolutionary language and to act in a reformist manner.
Einige waren für Revolution, andere für Reform, die meisten zogen es vor, eine revolutionäre **Sprache** zu sprechen, aber reformistisch zu handeln.

Abbildung 6: German/English-Translation-Korpus: Konkordanz für *Sprache* (Ausschnitt)

Konkordanzen, so kann man abschließend festhalten, dienen also dazu, Kontextinformationen zugänglich zu machen. Sie liefern jedoch keine Interpretation. Diese vorzunehmen ist die Aufgabe des Korpuslinguisten.

Aufgabe 14: Bitte erstellen Sie eine Konkordanz für das Wort *Absatz*. Benutzen Sie dazu ein beliebiges Korpus mit Konkordanzfunktion wie das DWDS-Korpus oder ein Programm wie Cosmas II oder WebConc.
Überprüfen Sie anschließend 50 Treffer. Welche Bedeutungsvarianten für *Absatz* finden Sie?

3.6 Kollokationsanalyse – die Suche nach benachbarten Wörtern

Um einen Überblick über den Kontext zu erhalten, in dem ein Wort steht, können Konkordanzen nach ihrem linken oder rechten Kontext sortiert werden. Abbildung 7 zeigt nochmals die Konkordanz für *Absatz* aus Abbildung 5, dieses Mal sind die einzelnen Zeilen jedoch nach dem rechten Kontext rückläufig sortiert. Konkret bedeutet dies, dass bei der Sortierung zuerst der letzte, dann der vorletzte, dann der drittletzte Buchstabe des rechten Kontextes berücksichtigt wird und so fort.

ndung von Wissenschaft , Produktion und **Absatz** in diesen starken ökonomischen Ei
r Forschung über die Produktion bis zum **Absatz** - weltmarktfähige Erzeugnisse her
Verteidigungswaffen fänden " reißenden **Absatz** " , berichtete das Hallesche Boule
die Schuhfabrik hat keinen **Absatz** mehr , beim Glühlampenhersteller
fkleber fanden unter Trabifahrern guten **Absatz** .
ltog sich im Schatten des Artikels 20, **Absatz** 2 der Verfassung , der die Diskrim
wie Senatoren hat – bestimmt Artikel 51 **Absatz** 2 des Grundgesetzes : .

Abbildung 7: Konkordanz für *Absatz* (sortiert nach rechten Kontext)

Findet eine solche Sortierung statt, so fallen häufig nebeneinander stehende Wortverbindungen wie etwa *reißenden Absatz finden* in Abbildung 7 leicht ins Auge. Sind zwei oder mehrere Wörter überdurchschnittlich oft benachbart, spricht man von **Kollokationen** oder Kookurrenzen. Wörter, die typischerweise in Verbindung mit einem Zielwort auftreten, werden als Kollokationspartner bezeichnet. Kollokationspartner zu *Himmel* sind im DWDS-Korpus *blau* (*blauer Himmel*), *grau* (*grauer Himmel*) oder *Erde* (*Himmel und Erde*), Kollokationspartner zu *blau* sind hingegen *rot*, *grün* oder *Himmel* (vgl. Kapitel 5.2).

Nach Kollokationspartnern wird jedoch häufig nicht innerhalb des gesamten Textes gesucht, sondern nur innerhalb einer festgelegten Textspanne. Wörter, die zwar häufig zusammen auftreten, aber weiter voneinander entfernt stehen, werden somit nicht mehr als Kollokationen erfasst. Anhand der Anzahl des gemeinsamen Auftretens von Zielwort und Kollokationspartner kann die Stärke einer Kollokation bestimmt werden. Kollokationen, deren Vorkommen deutlich die Wahrscheinlichkeit eines zufälligen Zusammentreffens übersteigen, werden als signifikante Kollokationen bezeichnet.

Für das Wort *Hund* hat Steyer (2002) auf der Grundlage der IDS-Textkorpora eine detaillierte Kollokationsanalyse vorgenommen. Ihr Ziel war es herauszufinden, ob sich mithilfe von Textkorpora sprachliches Wissen über bestimmte Begriffe, deren Bedeutung und Verwendung rekonstruieren lässt, das über die üblichen Wörterbuchinhalte hinausgeht. Als typische Kollokationspartner von *Hund* fand Steyer unter anderem die in (17) genannten Wörter *Leine*, *bellen* usw.

(17) Leine, bellen, Herrchen, Rassen, beißen, Schwanz, wedelt, Gassi, Haustiere, Zucht, streicheln

Diese Kollokationen von *Hund* stehen in Einklang mit unserem kulturellen Wissen über Hunde: Hunde sind Haustiere, sie werden von ihrem Herrchen an der Leine Gassi geführt, sie bellen, beißen, wedeln mit dem Schwanz, sie lassen sich streicheln, Hunde gehören zu verschiedenen Rassen, die gezüchtet werden, und so fort. Wie die Kollokationspartner von *Hund* in (17) zeigen, lassen sich Korpora also gut nutzen, um stereotypes Wissen über bestimmte Begriffe zu ermitteln.

Steyers Kollokationsanalyse zeigte darüber hinaus aber auch, dass das Wort *Hund* häufig in Zusammenhang mit Wörtern aus dem Wortfeld Familie wie *Vater*, *Mutter*, *Kind*, *Oma* oder *Haus* auftritt. Dies deutet darauf hin, dass *Hund* im Deutschen in ein bestimmtes Stereotyp von Familie eingebettet ist. Schließlich untersuchte Steyer mithilfe der Kollokationsanalyse das Auftreten von *Hund* in idiomatischen Wendungen. Dabei fand sie nicht nur bekannte Sprichwörter wie in (18), sondern stellte zusätzlich neue Mehrwortverbindungen wie in (19) fest. Diese neuen Ausdrücke haben in der Umgangssprache bereits den Charakter von idiomatischen