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*