

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 Suchwort 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. Teilentbach (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 Bei-

spiel 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 *REGGE*.

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. Homographie 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 *REGGE* zu unterscheiden (vgl. Kapitel 5.3).

Homographie 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 in Sonntag mit dem Bus 721 zum Flughafen gefahren. Die Haltestelle Bankstraße ist direkt bei kleinen einen vollbeladenen Möbelwagen und fahren davon. Aus einer kleinen Probefahrt wird in ich gern mal mit BVG zum Einkäufen gefahren. Jetzt nehme ich mir noch Auto - ist auf ein stens einem freien Feld jetzt prima schwarz fahren. Oder habe ich was falsch verstanden? Trac rüber 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 zurück übermüdet sind Gurten Sie sich immer an Fahren Sie defensiv und vorausschauend Denken Sie emals von der Straße weg und ins Gelände fahren. Aber man fährt 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 interessiert 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 Hallische Boule klieber fanden unter Trabfahrern guten Absatz  
 liegt sich im Schatten des Artikels 20, Absatz 2 der Verfassung , der die Diskriminierung Forschung über die Produktion bis zum Absatz - weltmarktfähige Erzeugnisse herndung 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 ausmacht. 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.

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 Eit Forschung über die Produktion bis zum Absatz - weltmarktfähige Erzeugnisse her Verteidigungswaffen fänden " reißenden Absatz ", berichtete das Hallesche Boule die Schlaufabrik hat keinen Absatz mehr , beim Glühlampenhersteller

flieber fanden unter Trabfahrern guten Absatz .

Itog 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.

net. 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 Zusammen treffens 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*, *bell*en usw.

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

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 stereotypen 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

Wendungen erreicht, sind bisher aber noch nicht in Wörterbüchern verzeichnet.

- (18) a. Hunde, die bellen, beißen nicht.  
b. Man soll schlafende Hunde nicht wecken.  
(19) a. der Schwanz wedelt mit dem Hund  
b. von allen Hunden gehetzt sein  
c. ein harter Hund sein

Zusammenfassend lässt sich sagen, dass die Analyse von Kollokationen ein guter Weg ist, um die Bedeutung bzw. Bedeutungsvarianten und die Verwendung eines Wortes zu ermitteln. Entsprechend wird die Analyse von Kollokationen in der Lexikografie, aber auch beim Übersetzen oder im Fremdsprachunterricht eingesetzt. Wie Steyers Analyse gezeigt hat, kann man Kollokationen aber auch dazu nutzen, um Stereotypen abzufragen, die mit einzelnen Ausdrücken verbunden sind.

**Aufgabe 15:** Erstellen Sie eine Konkordanz für die Wörter *Mann* und *Frau*. Benutzen Sie dazu ein beliebiges Korpus mit Konkordanzfunktion wie das DWDS-Korpus oder ein Programm wie Cosmas II oder WebConc. Überprüfen Sie anschließend je 50 Treffer. Welche Kollokationen finden Sie? Welche Eigenschaften werden Männern, welche Frauen zugeschrieben?

### 3.7 Wortlisten – Überblick über den Wortschatz

Eine einfache Möglichkeit, wenn man nicht nur einzelne Wörter, sondern den gesamten Wortschatz einer Sprache oder Varietät untersuchen will, ist es, ausgehend von einem entsprechenden Korpus eine **Wortliste** aller enthaltenen Textwörter zu erstellen. Diese Wortlisten ermöglichen es, einen groben Überblick über das in einem Korpus enthaltene sprachliche Material zu erhalten. Durch den Vergleich von Wortlisten lassen sich z.B. Fach- und Spezialwortschätze von der Standardsprache abgrenzen. Als medizinische Fachwörter wären etwa all jene Wörter aus einem medizinischen Textkorpus einzuordnen, die auf der Wortliste eines standardsprachlichen Referenzkorpus fehlen.

Überlegt man sich, welche Wörter im Deutschen besonders häufig sind, wird man vermutlich spontan an Wörter wie in (20) denken, an Wörter also, die Lebewesen, Gegenstände und Tätigkeiten bezeichnen, mit denen jeder von uns tagtäglich zu tun hat, und deren Entsprechungen man im Fremdsprachunterricht zuerst lernt.

(20) Mann, Frau, Kind, Haus, Auto, essen, trinken, schlafen

Umso erstaunter ist man, wenn man sich korpusbasierte Wortfrequenzlisten ansieht. Die folgende Wortliste in Abbildung 8 zeigt die fünfzig häufigsten Wortformen aus den Textkorpora des IDS. Zu beachten ist, dass die Mannheimer Wortliste orthografische Wortformen auflistet. Dies führt dazu, dass die Artikel *DER* und *DAS* jeweils in einer klein und einer groß geschriebenen Variante auftreten.

1	der	11	für	21	als	31	nach	41	vor
2	die	12	im	22	auch	32	am	42	einem
3	und	13	ist	23	es	33	bei	43	über
4	in	14	auf	24	an	34	wird	44	Das
5	den	15	des	25	aus	35	einer	45	einen
6	von	16	nicht	26	sie	36	um	46	zum
7	zu	17	Die	27	werden	37	wie	47	nur
8	mit	18	dem	28	er	38	daß	48	war
9	das	19	ein	29	hat	39	sind	49	so
10	sich	20	eine	30	Der	40	noch	50	haben

Abbildung 8: Die 50 häufigsten Wortformen in den Textkorpora des IDS

Es fällt auf, dass sich auf dieser Wortliste weder das Wort *Mann*, noch *Frau*, noch *Kind*, noch eines der anderen erwarteten Wörter findet – im Gegenteil: Unter den häufigsten 50 Wortformen ist kein einziges Nomen, kein Adjektiv und, was Verben anbelangt, so findet man lediglich Wortformen der Hilfsverben *haben*, *sein* und *werden*. Die häufigsten Formen von Modalverben, Vollverben, Nomen und Adjektiven sind in (21a-d) aufgeführt. In Klammern wird der jeweilige Rang in der Wortliste des IDS angegeben.

- (21) a. kann (68), soll (83), können (84), will (93)  
b. gibt (104), sagte (113), sagt (130)  
c. Uhr (56), Mark (74), Prozent (77), Jahren (78), Jahr (78), Jahre (95)  
d. neue (116), neuen (121), viel (191)

Der Vollständigkeit halber sei noch erwähnt, auf welchen Rängen sich die vermeintlich häufigsten Wörter aus (20) finden:

- (22) Mann (216), Frau (205), Kind (808), Haus (364), Auto (842), essen (4005), trinken (5240), schlafen (6767)

**Aufgabe 16:** Bitte vergleichen Sie die Wortliste des IDS mit der Wortliste des Projekts Deutscher Wortschatz in Leipzig. Zu welchem Ergebnis kommen Sie?

Welcher Schluss lässt sich aus dieser Wortliste ziehen? Zum einen zeigt die Wortliste deutlich, dass Funktionswörter wie Präpositionen, Artikel, Pronomen oder Hilfsverben die häufigsten Wortfor-

men in einer Sprache darstellen. Zum anderen sieht man, dass es sich bei den häufigsten Nomen und Adjektiven um relativ allgemeine Begriffe handelt, bei den Nomen etwa um Maßeinheiten, die Zeit, Geld oder Mengen beschreiben. Funktionswörter führen typischerweise die Frequenzlisten an – und das unabhängig von der Textsorte. Da Funktionswörter den Kitt bilden, der einen Text grammatisch zusammenhält, finden sie sich in jedem Korpus in großer Zahl, wohingegen die Anzahl und Art der Inhaltswörter wie Nomen, Adjektive oder Verben stärker vom Inhalt und der Textsorte abhängt.

Wortlisten können nach unterschiedlichen Kriterien gebildet und geordnet werden. Zum einen gibt es Wortlisten, die wie die Wortliste des IDS Wortform-Typen auflistet, andere Wortlisten wiederum fassen die einzelnen Wortformen zu Lemma-Typen zusammen. Beide Arten von Listen können alphabetisch, nach Frequenz, nach Wortlänge oder nach dem ersten Auftreten eines Lemmas sortiert werden. Rückläufige Wortlisten, d.h. Wortlisten, die Wörter nach ihrem Wortende ordnen, können dazu dienen, charakteristische Wortendungen wie z.B. *-el* oder *-itis* zu identifizieren.

**Aufgabe 17:** Bitte erstellen Sie eine alphabetisch geordnete Frequenzliste (Wortform-Typen) für den Textausschnitt aus dem Mainzer Zeitungskorpus (vgl. Abbildung 3, Kapitel 3.1).

Enthalten Wortlisten nicht nur Wortform- oder Lemma-Typen, sondern auch Angaben zu deren Häufigkeit, so sind neben einer qualitativen Analyse des Wortschatzes (Welche Wörter finden sich überhaupt im Korpus?) auch quantitative Aussagen möglich (Wie häufig sind die einzelnen Wörter? Welche Wörter kommen besonders häufig vor?). Wortlisten mit Angaben zur Frequenz einzelner Lemma- oder Wortform-Typen lassen sich vielfältig nutzen.

Frequenzangaben können z.B. dabei helfen, die Schlag- und Schlüsselwörter eines Korpus bzw. der darin enthaltenen Texte zu identifizieren, da diese im Text überdurchschnittlich häufig auftreten. Dies ist insbesondere dann von Interesse, wenn es wie bei der Analyse von Wahlkampfreden oder Werbetexten darum geht, sprachliche Mittel herauszuarbeiten, die der Verfasser einsetzt, um die Adressaten in seinem Sinne zu lenken. Wortlisten mit den häufigsten 500 oder 1.000 Wörtern einer Sprache erlauben es, den Kernwortschatz einer Sprache oder Varietät zu bestimmen. Die so entstandenen Wortlisten können wie der Grundwortschatz Deutsch von Pfeffer (1970) im Fremdsprachenunterricht eingesetzt werden, um den Lernenden zuerst jene Wörter beizubringen, denen sie mit

der größten Wahrscheinlichkeit begegnen werden (vgl. Kapitel 1.5). Dass die Intuition eines Sprechers nicht unbedingt zu zuverlässigen Frequenzinschätzungen führt, zeigte die Wortliste des IDS.

Wortlisten auf der Grundlage von Wortform-Typen können zwar relativ leicht erstellt werden, bringen jedoch eine Reihe von Problemen mit sich. So unterscheiden Wortform-basierte Wortlisten wie die des IDS nicht zwischen homographen Wortformen (vgl. das Beispiel *Regen* in Kapitel 3.4). Auch die unterschiedlichen Bedeutungsvarianten eines polysemen Wortes wie *Absatz* können in Wortlisten nicht voneinander getrennt werden (vgl. Kapitel 3.5).

### 3.8 Zusammenfassung

Korpuslinguistische Beschreibungsebenen wie Textwort, Type und Token müssen von sprachlichen Beschreibungsebenen wie Lexem und Wortform unterschieden werden.

Korpusdaten können sowohl qualitativ als auch quantitativ ausgewertet werden. Beim Vergleich von Daten muss die Korpusgröße berücksichtigt werden.

Verbreitete Verfahren zur Analyse von Korpusdaten sind die Stichwortsuche, das Erstellen von Konkordanzen und Wortlisten sowie die Kollokationsanalyse.

**Grundbegriffe:** Frequenz, Hapax Legomenon, Kollokation, Konkordanz, KWIC, Normalisierung, Produktivität, qualitative Korpusanalyse, quantitative Korpusanalyse, Textwort, Token, Type, Type-Token-Verhältnis, Wortliste

#### Weiterführende Literatur

Einen knappen Überblick über methodische Fragen bieten Biber *et al.* (1998, Teil IV). Ausführlich mit der Korpusanalyse befasst sich Kennedy (1998, Kapitel 4), speziell mit quantitativen Analysen McEnergy/Wilson (2003, Kapitel 3). Kollokationen thematisieren Steyer (2002) und Hausmann (2004), Konkordanzen und Kollokationen Sinclair (1991). Mit der Abgrenzung des Wortes in der Linguistik bzw. Korpuslinguistik befassen sich Meibauer *et al.* (2002, Kapitel 2) und Haß-Zumkehr (2002). Nähere Informationen zu statistischen Verfahren bietet Schlobinski (1996).