Lothar Lemnitzer / Heike Zinsmeister

Korpuslinguistik

Eine Einführung

3. Auflage

narr STUDIENBÜCHER

narr francke attempto

Lothar Lemnitzer / Heike Zinsmeister

Korpuslinguistik

Eine Einführung

3., überarbeitete und erweiterte Auflage

narr\f
ranck
e\atte
mpto

Dr. Lothar Lemnitzer ist Wissenschaftlicher Mitarbeiter am *Digitalen Wörterbuch Sprache des 20. Jahrhunderts* (*DWDS*) an der Berlin-Brandenburgischen Akademie der Wissenschaften.

Prof. Dr. Heike Zinsmeister lehrt Linguistik des Deutschen und Korpuslinguistik am Institut für Germanistik der Universität Hamburg.

3., überarbeitete und erweiterte Auflage 2015
2., durchgesehene und aktualisierte Auflage 2010
1. Auflage 2006

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Gedruckt auf säurefreiem und alterungsbeständigem Werkdruckpapier.

Internet: www.narr-studienbuecher.de
E-Mail: info@narr.de

Printed in the EU

ISSN 0941-8105
ISBN 978-3-8233-6886-1

Inhalt

Dr. Lothar Lemnitzer ist Wissenschaftlicher Mitarbeiter am *Digitalen Wörterbuch Sprache des 20. Jahrhunderts* (*DWDS*) an der Berlin-Brandenburgischen Akademie der Wissenschaften.

Prof. Dr. Heike Zinsmeister lehrt Linguistik des Deutschen und Korpuslinguistik am Institut für Germanistik der Universität Hamburg.

3., überarbeitete und erweiterte Auflage 2015
2., durchgesehene und aktualisierte Auflage 2010
1. Auflage 2006

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Gedruckt auf säurefreiem und alterungsbeständigem Werkdruckpapier.

Internet: www.narr-studienbuecher.de
E-Mail: info@narr.de

Printed in the EU

ISSN 0941-8105
ISBN 978-3-8233-6886-1

Inhalt

Vorwort zur 3. Auflage

Durch den Wechsel auf das moderne Layout der Reihe *narr Studienbücher* konnten wir die alten Inhalte kompakter darstellen. Zusätzlich bedanken wir uns beim Verlag für die Option, die Gesamtseitenzahl etwas zu erhöhen, sodass Sie nun eine gründlich durchgesehene, aktualisierte und erweiterte Version unseres Buches in Händen halten.

In das zweite Kapitel zu den theoretischen Grundlagen der Korpuslinguistik haben wir die Arbeiten von András Kertesz und Csilla Rákosi aufgenommen. Die Arbeiten der beiden Autoren scheinen uns der interessanteste Beitrag der letzten Jahre zur wissenschaftstheoretischen Begründung der Korpuslinguistik zu sein.

Im dritten Kapitel haben wir neuere Entwicklungen bei den Standards für Metadaten berücksichtigt, die sich aus der Arbeit sprachressourcen-bezogener Projekte wie CLARIN ergeben haben.

Kapitel 4 wurde grundlegend überarbeitet und um ein Unterkapitel zur manuellen Annotation und der Entwicklung von Annotationsschemata erweitert, sodass wir uns entschlossen, die Inhalte in der neuen Auflage auf zwei Kapitel zu verteilen. Das neue Kapitel 5 endet mit vier praktischen Übungen zur Suche auf Online-Korpora.

Für die vorliegende Auflage haben wir eine Anregung von Markus Hundt (2006) aus seiner Rezension zu unserer Erstauflage aufgegriffen und ein zusätzliches Kapitel zur quantitativen Auswertung von Korpusdaten ergänzt. Das neue Kapitel 6 hat daher keine Entsprechung in den alten Auflagen. Es soll die Leser dafür sensibilisieren, wie linguistische Fragestellungen „korpustauglich" gemacht und in eine quantitativ überprüfbare Hypothese überführt werden können. Darüber hinaus führt das Kapitel in die deskriptive Statistik ein, in der Verteilungen durch statistische Kennwerte beschrieben und grafisch dargestellt werden.

Kapitel 7 (ehemals Kapitel 5) ist im Vergleich mit den alten Auflagen geschrumpft, da wir die tabellarische Beschreibung von Einzelkorpora, die bisher ein Bestandteil dieses Kapitels war, komplett auf die begleitende Webseite ausgelagert haben. Dafür haben wir einige Abschnitte hinzugefügt, in denen neue Korpora und Korpusinitiativen beschrieben werden.

In Kapitel 8 (ehemals 6) wurden die Kategorien, unter die die Fallstudien eingeordnet sind, weitgehend beibehalten, aber anders angeordnet. Innerhalb der einzelnen Fallstudien wurden wichtige neuere Forschungsarbeiten ergänzt.

Das abschließende Kapitel (Erfahrungen von Linguistinnen mit der Verwendung von Korpora) wurde komplett auf die Webseite ausgelagert. Es konnte im Rahmen dieser Überarbeitung nicht aktualisiert werden und hat in unseren Augen eher historisch-dokumentarischen Wert. Wir empfehlen stattdessen die Lektüre des Artikels von Detmar Meurers und Stefan Müller (2008), der viele hilfreiche Erkenntnisse liefert.

Als Services stehen weiterhin das aktualisierte Literaturverzeichnis, das Glossar und das ebenfalls aktualisierte Sachregister zur Verfügung. Ferner wurden wichtige Textstellen mit Marginalsymbolen versehen: Ein Buch markiert weiterführende Literatur, ein Bleistift Aufgaben, ein Ausrufezeichen Hinweise und eine Gedankenblase Anregungen zum Weiterdenken. Die buchbegleitende Webseite finden Sie unter `www.narr-studienbuecher.de/9783823378860`. Wir bedanken uns ganz herzlich bei unserem neuen Lektor Tillmann Bub für seine beharrliche und sehr konstruktive Betreuung der Neuauflage.

Berlin und Hamburg, im September 2015 Lothar Lemnitzer & Heike Zinsmeister

Vorwort zur 1. Auflage

Im Frühjahr 2005 wurden wir gefragt, ob wir eine Einführung in die Korpuslinguistik für Germanisten schreiben wollten. Wir stellten uns dieser Aufgabe gerne, da es bis jetzt kein deutsches Lehrwerk für die korpuslinguistische Lehre oder für das Selbststudium gibt. Andererseits zeigt die große Zahl an korpuslinguistischen Seminaren in der Germanistik und allgemeinen Sprachwissenschaft, dass Bedarf an einem Lehrwerk besteht.

Bei der Recherche für dieses Thema waren wir überrascht, wie viele korpuslinguistische Untersuchungen mit einem weiten thematischen Spektrum mittlerweile veröffentlicht wurden. Es war eine Freude, diese Arbeiten mit einer korpuslinguistischen Brille zu lesen und auszuwerten. Wir sind sicher, dass auch Sie als Leser von dieser Zusammenschau profitieren werden. Wenn Sie sich durch dieses Buch zu eigener korpuslinguistischer Arbeit ermutigt fühlen, dann haben wir unser wichtigstes Ziel erreicht.

Wir nutzen die Gelegenheit, um uns bei unserem Lektor Jürgen Freudl für die Anregung zu diesem Buch und für die Unterstützung bei unserer Arbeit zu bedanken. Dank gebührt auch den Testlesern der Vorversionen dieses Buches: Karin Pittner und Judith Berman haben eine Vorversion des Buches in ihrem Seminar getestet; Stefanie Dipper, Stefan Engelberg, Michael Götze, Anke Lüdeling, Sabine Schulte im Walde und Elke Zinsmeister haben wertvolle Kommentare zu einzelnen Kapiteln gegeben. Die verbleibenden Fehler gehen natürlich auf unsere Kappe.

Wir danken allen Kolleginnen und Kollegen, die sich spontan zu einem Interview oder einer schriftlichen Stellungnahme zu unseren Fragen bereit erklärt haben. Das Ergebnis können Sie in Kapitel 7 nachlesen.

Unser Dank gilt natürlich auch unseren Familien, Freunden, Kolleginnen und Kollegen, die unser eigentümliches Verhalten vor allem in der Abschlussphase dieses Buches mit Geduld ertragen haben. Ohne ihre Unterstützung wäre dieses Buch nicht das geworden, was es ist.

Schließlich möchten wir Ihnen danken, wenn Sie dieses Buch käuflich erworben haben. Wir freuen uns auf Ihre kritische Begleitung und auf Ihre Kommentare. Schreiben Sie uns! Unsere Adressen finden Sie auf der begleitenden Webseite.

Tübingen, im Februar 2006 — Lothar Lemnitzer & Heike Zinsmeister

Zum Geleit

Bis vor einigen Jahren schien es fast so, also wolle die germanistische Linguistik in Deutschland die Möglichkeiten der Korpuslinguistik verschlafen. Es gab zwar einige computerlinguistische Zentren, die mit zum Teil sehr großen Korpora arbeiteten, und auch die Korpora des IDS in Mannheim, aber korpuslinguistische Methoden wurden an den germanistischen Instituten an den Universitäten kaum unterrichtet und zum Teil immer noch kritisch beäugt. Das hat sich in den letzten Jahren gründlich geändert. An vielen Stellen gibt es inzwischen korpuslinguistische Seminare, Projekte und Sonderforschungsbereiche.

Dabei hat sich der alte Streit zwischen der Theorie und der Empirie längst entschärft und zu einem konstruktiven Miteinander gewandelt. Wir haben verstanden, dass unterschiedliche Fragestellungen auch unterschiedliche Daten erfordern und dass wir gemeinsame Ressourcen, Verfahren und Standards brauchen und diese deshalb entwickeln und evaluieren müssen. Für viele Fragestellungen, zum Beispiel zu historischen Untersuchungen oder zu Erwerbsprozessen im Erst- und Zweitspracherwerb, liegen schlicht keine anderen Daten vor, gerade hier ist man auf allgemein zugängliche und standardisierte Ressourcen und Werkzeuge angewiesen.

Eine Grundlage für gute korpusbasierte Arbeit ist gute korpuslinguistische Lehre. Es gibt eine Reihe von englischsprachigen Einführungsbüchern, die sich mit Korpora beschäftigen – diese konzentrieren sich jedoch auf englischsprachige Ressourcen und Studien. Bisher fehlte ein korpuslinguistisches Einführungsbuch für Germanistikstudierende ohne informatische Vorkenntnisse, das sich speziell auf die deutschen Korpora und Fragestellungen bezieht. Lothar Lemnitzer und Heike Zinsmeister haben nun eine solche interessante, fundierte und klar geschriebene Einführung in die Korpuslinguistik vorgelegt. Das Buch beschäftigt sich zunächst mit den linguistischen Grundlagen der Korpuslinguistik und lotet dabei die Chancen und Grenzen der Arbeit mit Korpora aus. Zusätzlich werden Methoden der Datengewinnung und Annotation erläutert und diskutiert. Konkrete Studien aus ganz unterschiedlichen linguistischen Bereichen zeigen anschaulich, wie breit korpuslinguistische Verfahren in der linguistischen Forschung eingesetzt werden können. Ich freue mich darauf, mit diesem Buch arbeiten zu können.

Berlin, im Februar 2006

Anke Lüdeling
Professorin für Korpuslinguistik
Humboldt-Universität zu Berlin

1 Einleitung

1.1 Was ist Korpuslinguistik?

Die Folklore der Sprachwissenschaft[1] kennt zwei Forschertypen:

- **Der Denker**[2] verbringt die meiste Zeit in seinem Sessel und denkt nach. Die Sprachtheorie, die er sich mit den Jahren in seinem Kopf zurechtgelegt hat, wird durch Beispiele, die unmittelbar seiner Sprachkompetenz entspringen, bestätigt oder widerlegt. Hin und wieder notiert sich der Denker besonders komplizierte und abwegige Beispiele, die durch eine Grammatik, die dieser Sprachtheorie entspricht, hergeleitet werden können. Diese Sätze legt er Sprechern der untersuchten Sprache mit der Frage vor, ob diese Sätze denn wohlgeformt seien. Daraus, ob die befragten kompetenten Sprecher seine Beispiele gutheißen oder ablehnen, zieht der Denker weit reichende Schlüsse über den Aufbau der Grammatik dieser Sprache und der zugrunde liegenden Sprachtheorie. Was für den Denker alleine zählt, ist das Urteil kompetenter Sprecher, das auf deren Sprachgefühl und sprachlichem Wissen fußt. Der Denker hält sich an den Rändern der Sprache auf, in Bereichen, die wenig mit dem alltäglichen Sprachgebrauch zu tun haben. Im Gegenteil, der Denker ist an den Äußerungen, die tagtäglich produziert werden, herzlich wenig interessiert. Sie sind wenig erleuchtend für seine Theorie.
- **Der Beobachter** ist an authentischen Sprachdaten interessiert: je mehr Daten, desto besser. Die Theorien, die er entwickelt, sind auf die Beobachtung dieser Daten gestützt. Seine Aussagen und Hypothesen werden durch immer neue Daten bestätigt oder verworfen. Mit seinen Kollegen spricht der Beobachter vor allem darüber, welche interessanten Beobachtungen er gemacht hat. Ansonsten hält er sich überwiegend an seinem Computer auf. Das Bild, das er durch diese Beobachtungen gewinnen möchte, sollte möglichst vollständig sein, deshalb ist er vor allem an den Phänomenen interessiert, die in unserem alltäglichen Sprachgebrauch vorkommen.

---

[1] Wer nicht glaubt, dass es eine Folklore der Sprachwissenschaft gibt, der möge sich einmal Pullum (1991) ansehen. Auch allen anderen Lesern möchten wir dieses vergnüglich zu lesende Buch empfehlen.

[2] Wir verwenden in diesem Buch das generische Maskulinum bei Bezeichnungen von Personen und schließen damit selbstverständlich alle weiblichen Personen mit ein. Die Wahl dieser Form hat einzig und allein den Grund, dass ihre Verwendung das Lesen des Textes etwas einfacher macht.

Der Denker erweist sich als scharfsinniger Theoretiker, der die Grundlagen des Sprachvermögens erforscht, das allen Menschen gemeinsam ist, und dies *Universalgrammatik* nennt. Für seine Forschungen muss er seinen Sessel nur äußerst selten verlassen. Den Beobachter hingegen findet man häufig dort, wo es um die möglichst umfassende Beschreibung einer Sprache in ihrer alltäglichen Verwendung und die Vermittlung dieses Sprachgebrauchs, z.B. in Lexikographie und Fremdsprachunterricht, geht.

Diese plastische Beschreibung zweier Typen von Forschern in der Linguistik ist nicht neu. Sie findet sich so ähnlich schon bei Charles Fillmore (Fillmore, 1992). Fillmore hat in den achtziger Jahren das Lager gewechselt und sich vom theoretisierenden Linguisten zum Beobachter gewandelt. Es ist jedoch keinesfalls so, dass die Entscheidung für eine Richtung die andere Richtung ausschließt: Wer sammelt, hat damit das Denken nicht aufgegeben, und auch der Denker profitiert hin und wieder von den Erkenntnissen der Beobachter. Wir werden Beispiele dafür noch kennen lernen.

Eine Einführung in die Korpuslinguistik wendet sich in erster Linie an die Beobachter unter den Sprachwissenschaftlern. Wer Korpuslinguistik betreibt, dem geht es in erster Linie um das Beobachten und Beschreiben sprachlicher Phänomene. Wir wenden uns aber auch an die Denker und werden zeigen, dass und wie sie von den Beobachtungen und Erkenntnissen der Korpuslinguisten profitieren können. Eine enge Zusammenarbeit zwischen Denkern und Beobachtern, also zwischen theoretischen Linguisten und empirisch arbeitenden Linguisten, erscheint uns fruchtbar für beide Seiten. Eine solche Haltung ist in der Zunft aber keinesfalls selbstverständlich. Randy Allen Harris hat sein Buch über die Sprachwissenschaft in den sechziger und siebziger Jahren des letzten Jahrhunderts „Linguistic Wars" genannt, und dies ist sicher nicht allzu stark übertrieben. Charles Hockett, ein Vertreter der empirischen Arbeitsweise, bezeichnete die Methode, Selbstauskünfte von Sprechern über ihr sprachliches Wissen heranzuziehen, als im günstigsten Fall überflüssig (*superfluous*) und im ungünstigsten Fall als widerwärtig (*obnoxious*)[3]. Viele theoretische Sprachwissenschaftler im Umfeld der generativen Sprachtheorie, allen voran Noam Chomsky, bezeichnen das Werk der Korpuslinguistik als irrelevant und nutzlos[4]. Es gibt, wie gesagt, Berichte von „Lagerwechseln"[5], was auch nicht gerade für ein friedliches Zusammenleben spricht.

Wir werden im zweiten Kapitel zeigen, dass mindestens ein Teil der Kritik, die von Sprachtheoretikern gegenüber empirisch arbeitenden Linguisten geäußert wurde, berechtigt ist. Sie betrifft Annahmen, die von der Korpuslinguistik in der Zeit vor dem Entstehen der generativen Grammatik in den fünfziger Jahren getroffen wurden. Die moderne Korpuslinguistik hat daraus gelernt. Es ist aber auch heute noch so, dass jeder, der korpuslinguistisch arbeitet, eine Antwort auf die Kritik aus dem sprachtheoretischen Lager haben sollte. Wir werden auf diese Antworten ausführlicher im dritten Kapitel eingehen.

Zunächst jedoch wollen wir eine Antwort auf die Frage geben, was Korpuslinguistik eigentlich ist. Das Wort ist ein Kompositum, es setzt sich aus den Bestandteilen *Korpus*

---

3 Vgl. Hocket (1964), zitiert nach McEnery und Wilson (1996). Wir werden in Abschnitt 2.2 auf die Probleme eingehen, die Selbstauskünfte von Sprechern tatsächlich mit sich bringen.

4 Z.B. Chomsky (1986), S. 27.

5 Vgl. zum Beispiel Fillmore (1992) und Sampson (1996).

und *Linguistik* zusammen. Eine Antwort auf die Frage führt also zunächst über diese beiden Begriffe.

Definition 1 (Korpus[6]). *Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind*[7].

Die Sammlung von Äußerungen ist meist das Ergebnis sorgfältiger Planung[8], was nicht ausschließt, dass auch ad hoc oder zu anderen als linguistischen Zwecken entstandene Textsammlungen einen Wert als Datenbasis haben können. Je besser ein Korpus geplant ist, um so nützlicher ist es für die spätere Forschung.

Heutzutage liegen Korpusdaten in maschinenlesbarer Form vor. Es gibt auch heute noch nicht digitalisierte Textsammlungen bzw. Recherchen, die sich auf solche beziehen. Wir werden in Kapitel 8 solche Untersuchungen vorstellen. Die Verwendung nicht digitalisierter Texte führt jedoch zu methodischen Problemen. Auch dies werden wir in Kapitel 8 zeigen. „Ältere Texte werden heute in vielen Projekten nachträglich digitalisiert. Das Gleiche gilt für Tonaufzeichnungen von Interviews, Gesprächen usw. Man tut gut daran, sich Gedanken zu machen, ob es digitalisierte Daten für die eigenen Untersuchungen gibt bzw. ob und wie man die eigenen Daten digitalisieren kann. Wir betrachten hier das digitale Korpus als die Norm.

Der Wert eines Korpus wächst, wenn seine *Primärdaten* mit beschreibenden Daten versehen werden, die z.B. Auskunft geben über die Autoren von Texten oder die Sprecher von Tonaufnahmen, über den Zeitpunkt der Entstehung usw. Man spricht hierbei auch von *Metadaten*. Von diesen Daten, die ganze Texte oder zusammenhängende Äußerungsfolgen beschreiben, unterscheiden wir die *Annotationen*, die sich auf Teile von Texten bezieht, also auf Wörter, Sätze, Paragraphen usw. Annotationen markieren und klassifizieren bestimmte Einheiten, zum Beispiel Wörter mit ihrer Wortart.

Von anderen Medien außer Text oder Ton sehen wir ab, wollen aber darauf hinweisen, dass es interessante Korpora gibt, in denen Text und Ton mit stehenden oder bewegten Bildern verbunden werden. Man spricht dann von *multimedialen* oder *multimodalen* Korpora[9].

Der zweite konstituierende Begriff ist *Linguistik*. Diese Disziplin wird im deutschen Sprachraum meistens als *Sprachwissenschaft* bezeichnet. Damit ist der Gegenstand dieser Disziplin im weitesten Sinn umschrieben. Das Wort *Sprache* ist aber mehrdeutig, wie die folgenden Beispiele zeigen:

---

6 Im Deutschen wird das Neutrum verwendet, es heißt also *das Korpus*, wenn von einer Sammlung von Äußerungen die Rede ist. In allen anderen Bedeutungen wird das Wort im Maskulinum verwendet.

7 In diesem Buch wird es überwiegend um Korpora geschriebener Texte gehen. Eine gute Einführung in Korpora gesprochener Sprache liegt nun mit Draxler (2008) vor.

8 Vgl. hierzu ausführlich Hunston (2008).

9 Einen guten Überblick über multimodale Korpora gibt Jens Allwood (2008).

(1) ...weil Deutsch die Sprache ist, in der ich meine Gedanken am schönsten darlegen kann. (taz, 25.6.1993)

(2) ...als ich die ersten Bilder sah, verschlug es mir die Sprache. (taz, 15.11.1996)

(3) Aber auch der Kosovo, Afghanistan und der Kaukasus kamen zur Sprache. (taz, 5.2.1999)

(4) Sie verzichten darauf, Hölderlins Sprache mit Bedeutung aufzuladen. (taz, 6.8.1990)

In Beispiel (1) ist mit *Sprache* eine konkrete natürliche Sprache, zum Beispiel das Deutsche, gemeint. In Beispiel (2) geht es allgemeiner um das Sprachvermögen und den Zugang zu diesem, welcher bei dem entgeisterten Betrachter momentan blockiert ist. Er wäre weder in der Lage sich in Deutsch, noch in irgendeiner anderen Sprache zu äußern. In Beispiel (3) ist mit *zur Sprache kommen* ein konkretes sprachliches Ereignis gemeint. In Beispiel (4) schließlich bezieht sich der Autor auf die Eigensprache einer einzelnen Person.

Dass mit *Sprache* Unterschiedliches bezeichnet werden kann, hat Auswirkungen auf die Wissenschaft von der Sprache bzw. den Sprachen. All die in diesen Beispielen dargestellten Aspekte können Gegenstand der wissenschaftlichen Betrachtung sein. Ein Grund für den Streit zwischen den verschiedenen sprachwissenschaftlichen Lagern ist, dass der Gegenstand der eigenen wissenschaftlichen Betrachtung verabsolutiert wird und die anderen Gegenstände nicht der wissenschaftlichen Untersuchung wert befunden werden.

Korpuslinguisten haben es mit Sprache in dem Sinn zu tun, der in Beispiel (3) zum Ausdruck kommt. Die Korpora, die untersucht werden, stellen Sammlungen konkreter sprachlicher Äußerungen dar. Natürlich werden diese in einer bestimmten Sprache getätigt, z.B. im Deutschen, Spanischen oder Chinesischen. Wir werden uns in diesem Buch auf deutsche Korpora und die korpuslinguistische Untersuchung der deutschen Sprache konzentrieren[10]. Inwieweit von Äußerungen als Gegenstand der Untersuchung auf das Sprachvermögen der Sprecher geschlossen werden kann, ist umstritten. Es ist sogar umstritten, ob dies ein wissenschaftliches Ziel der Korpuslinguistik sein sollte[11].

Nach diesen Begriffsbestimmungen wollen wir nun versuchen, eine Antwort auf die Eingangsfrage zu geben: Was ist Korpuslinguistik?

Definition 2 (Korpuslinguistik). *Man bezeichnet als Korpuslinguistik die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind. Korpuslinguistik ist eine wissenschaftliche Disziplin, d.h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Korpusbasierte Sprachbeschreibung kann verschiedenen Zwecken dienen, zum Beispiel dem*

---

[10] Natürlich ist der Begriff *deutsche Sprache* selbst eine Abstraktion, die von Dialekten wie dem Schwäbischen, nationalen Varianten wie dem Österreichischen oder Fachsprachen wie der Sprache der Informatik abstrahiert. Von diesen Varietäten kann man zu Recht fragen, in wie weit diese noch *deutsche Sprache* sind. Das Konstrukt *deutsche Sprache* ist jedoch den meisten Sprechern vertraut und hat sich als übergeordneter Begriff auch in der Sprachwissenschaft bewährt.

[11] „...the task of corpus linguists is to exemplify the dominant structural patterns of the language without recourse to abstraction, or indeed to generalization" (Sinclair, 1991, S. 103).

Fremdsprachunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen Sprachverarbeitung bzw. Computerlinguistik.

Gegenstand von Korpora und damit der Korpuslinguistik sind natürliche Sprachen, nicht formale Sprachen wie z.B. Programmiersprachen. Das schließt die Untersuchung von älteren Sprachstadien natürlicher Sprachen, wie etwa des Althochdeutschen oder des Mittelhochdeutschen, ein. Eine Vorbedingung ist allerdings, dass die überlieferten Texte dieser Sprachdenkmäler in digitalisierter Form vorliegen. In den letzten Jahren werden solche Texte in verstärktem Maße digitalisiert, man spricht dabei von *Retrodigitalisierung*[12]. Eine Stärke der Korpuslinguistik ist es, dass auf Grund der Datenbasis nicht nur die Struktur einer Sprache, sondern auch deren Verwendung untersucht werden kann.

Die Einhaltung gewisser Prinzipien ist die Grundvoraussetzung jeder wissenschaftlichen Tätigkeit. Dazu gehört, dass die Ergebnisse von Untersuchungen nachprüfbar oder sogar reproduzierbar sein müssen. Im Falle der Korpuslinguistik bedeutet dies, dass die Ergebnisse von Untersuchungen durch andere Personen am selben Korpus nachvollziehbar sein sollten. Idealerweise sollte es zusätzlich möglich sein, die Untersuchungen auch an vergleichbaren, anderen Korpora als denen, auf die sie sich ursprünglich stützen, nachzuprüfen. Die gemeinsame Nutzung eines Korpus für verschiedene Untersuchungen gewährleistet, dass Forschungsergebnisse miteinander verglichen werden können. Die Methoden der Untersuchung sollten den anerkannten wissenschaftlichen Standards entsprechen, und es muss Klarheit bestehen über die Reichweite und Sicherheit von Aussagen, die auf Grund von Beobachtungen getroffen werden. Dies trifft gleichermaßen für statistische über Regularitäten wie für Gesetzesaussagen zu. Statistische Aussagen benennen Tendenzen in den Daten, die durch einzelne Gegenbeispiele nicht widerlegt werden können. Bei dieser Art von Aussagen sollte aber die Sicherheit angegeben werden können, mit der die Aussage zutrifft. Hierfür gibt es in der Statistik etablierte Verfahren. Gesetzesaussagen hingegen sind absoluter – sie bezeichnen Regeln und Zusammenhänge, die immer zutreffen. Deshalb sind sie leichter, nämlich bereits durch ein einziges Gegenbeispiel, widerlegbar.

Korpuslinguistik ist stärker als andere Richtungen der Sprachwissenschaft zweckorientiert. Die Erkenntnisse der Korpuslinguistik beeinflussen u.a. die Übersetzungswissenschaft, die Lexikografie und den Sprachunterricht.

1.2 Wer sollte dieses Buch lesen?

Diese Einführung wendet sich an Studierende und Forscher der Sprachwissenschaft, die empirisch die deutsche Sprache untersuchen wollen. Wir wollen ihnen mit diesem Buch das Wissen und die Mittel an die Hand geben, die für die Planung und Durchführung korpuslinguistischer Untersuchungen benötigt werden. Sie sollen mit diesem Buch in die Lage versetzt werden, ein für ihre Fragestellung geeignetes Korpus auszuwählen oder ein eigenes Korpus zu erstellen. Das Buch ist auch zum Selbststudium geeignet. Wir werden lediglich die Kenntnisse voraussetzen, die in einer allgemeinen Einführung in die (germanistische) Linguistik erworben werden können[13].

[12] Vgl. hierzu Altrichter (2001) und Claridge (2008).

[13] Zum Beispiel die allgemeine Einführung herausgegeben von Jacob Ossner und Heike Zinsmeister (2014) oder – für die syntaktische Analyse – das bewährte Arbeitsbuch von Karin Pittner und Judith Berman (2013).

1.3 Aufbau des Buchs

Im zweiten Kapitel werden wir ausführlicher auf die Kritik, die von sprachtheoretischer Seite gegen die Korpuslinguistik vorgebracht wurde, eingehen. Der Gegensatz zwischen Generativer Grammatik und Korpuslinguistik ist grundsätzlich. Er wurzelt in einer unterschiedlichen Auffassung von Gegenstand und Methode der Linguistik, wie wir darstellen werden. Wir stellen die im positiven wie negativen Sinne für die Korpuslinguistik einflussreichen linguistischen Strömungen der Generativen Grammatik und des Kontextualismus vor. Am Ende dieses Kapitels werden wir drei Ansätze korpuslinguistischer Forschung gegenüberstellen: einen korpusbasierten, rein quantitativen Ansatz, einen korpusbasierten, quantitativ wie auch qualitativ ausgerichteten Ansatz und einen korpusgestützten, qualitativen Ansatz.

Im dritten Kapitel werden wir ausführlicher darstellen, was linguistische Korpora sind, in Abgrenzung zu anderen Arten linguistischer Datensammlungen. Wir werden drei für linguistische Korpora relevante Datenebenen unterscheiden: die Primärdaten, die Metadaten und die linguistische Annotation. Für die Beschreibung linguistischer Korpora haben sich auf internationaler Ebene Standards durchgesetzt. Diese Standards werden wir vorstellen. Der abschließende Teil ist methodischen Problemen gewidmet, die man lösen sollte, bevor man Korpora für eine linguistische Untersuchung heranzieht. Wir werden die folgenden Fragen beantworten: Können Korpora repräsentativ sein? Wie findet man sprachliche Phänomene in großen Mengen von Sprachdaten? Was macht man, wenn ein zu untersuchendes Phänomen nicht im Korpus gefunden wird und was, wenn man etwas findet, das auf Grund einer entwickelten Theorie eigentlich nicht vorkommen dürfte?

Linguistische Annotationen helfen, relevante Vorkommnisse in größeren Datenmengen (wieder) zu finden. Hierzu werden die Daten linguistisch voranalysiert und mit Annotationen wie zum Beispiel Wortarten oder grammatischen Funktionen versehen. Wir werden im vierten Kapitel Mittel und Methoden der Annotation darstellen und dabei unter anderem syntaktisch annotierte Korpora vorstellen. Um die Leser an die eigene Nutzung solcher Ressourcen heranzuführen, werden wir im fünften Kapitel anschließend auf die linguistische Abfrage von Korpora eingehen, darüber hinaus Methoden für das eigene Annotieren und eine Reihe von Abfrage- und Annotierwerkzeuge vorstellen.

Im sechsten Kapitel werden wir die Leser an die quantitative Auswertung auf der Basis von Korpora heranführen. Dort, wo wir grundlegende Konzepte von Mathematik und Statistik benötigen, werden wir diese informell einführen und im Übrigen auf vertiefende Literatur zu diesem Thema hinweisen. Wir, die Autoren dieses Buches, haben die Erfahrung gemacht, dass es durchaus auch Nicht-Mathematikern gelingen kann, sich das Handwerkszeug quantitativer Forschung anzueignen.

Korpora unterscheiden sich in vielfältiger Weise. Im siebten Kapitel werden wir anhand von konkreten Korpusbeispielen eine Typologie einführen, in der verschiedene Korpustypen systematisch unterschieden werden. Am Schluss dieses Kapitels stellen wir einige vielversprechende neue Korpusinitiativen vor.

Korpora sind die Materialgrundlage vielfältiger qualitativer und quantitativer sprachwissenschaftlicher Untersuchungen. Im achten und letzten Kapitel werden wir einige ausgewählte Untersuchungen präsentieren und damit die Vielfalt der Fragen sichtbar machen, die mit Hilfe von Korpora beantwortet werden können.

Glossar und Index im Anhang werden sicherlich auch denen helfen, die das Buch zum Nachschlagen oder zum Lernen auf eine Prüfung verwenden wollen. Begleitet wird dieses Buch von einer Webseite, die unter `www.narr studienbuecher.de/9783823378860` erreichbar ist. Hier finden Sie:

- Eine tabellarische Liste von Korpusprojekten. Diese Einträge werden nach den in Kapitel 7 eingeführten Kriterien beschrieben;
- Hinweise auf Werkzeuge, die die Arbeit mit Korpora erleichtern;
- Handreichungen zu einigen der gebräuchlicheren Korpuswerkzeuge;
- Lösungsansätze für die Übungsaufgaben;
- weitere nützliche Links;
- weitere Informationen zu den Autoren des Buchs.

Diese Einleitung ist ein guter Ort, um über weitere Einleitungen und Handbücher zu informieren, die unsere Leser auch interessieren könnten. Eine weitere deutsche Einführung in das Thema hat Carmen Scherer verfasst (Scherer, 2005). Dieser Text kann als eine etwas leichtgewichtigere, an Germanisten gerichtete Alternative zu diesem Buch betrachtet werden. Drei Mitarbeiter des Instituts für Deutsche Sprache (IDS) haben eine methodisch ausgerichtete Einführung erarbeitet, die den Umgang mit großen Datenmengen und die speziellen Abfragemöglichkeiten der IDS-Korpora in den Mittelpunkt stellt (Perkuhn et al., 2012). Eine in Deutsch verfasste, aber an Anglisten gerichtete Einführung ist die von Joybrato Mukherjee (Mukherjee, 2009). Korpora gesprochener Sprache, die hier nur am Rande behandelt werden, stehen im Mittelpunkt einer gut lesbaren Einführung von Christoph Draxler (Draxler, 2008). Aus dem angelsächsischen Raum ist unbedingt das Buch von Tony McEnery, Richard Xiao und Yukio Tono zu erwähnen (McEnery et al., 2006). Es ist zum einen die Fortschreibung von McEnery und Wilson (2001), zum anderen enthält sie, über das ältere Werk hinausgehend, eine Dokumentation über wichtige methodische Diskussionen innerhalb der Korpuslinguistik (Teil B) und dreizehn Fallstudien, in denen beispielhaft Schritt für Schritt korpuslinguistische Projekte entwickelt werden (Teil C). Eine weiterführende Auseinandersetzung mit den verschiedenen Strömungen der Korpuslinguistik findet sich in McEnery und Hardie (2012). An Anfänger hingegen richtet sich die Einführung von Kübler und Zinsmeister (2015), die sich auf linguistisch annotierte Korpora konzentriert.

Neben diesen Einführungen sind auch zwei Handbücher erschienen. Bei de Gruyter wurden 2008 und 2009 zwei Bände des internationalen Handbuchs „Corpus Linguistics" veröffentlicht (Lüdeling und Kytö, 2008, 2009). Auf einige Aufsätze aus diesem Handbuch werden wir im Laufe dieses Buches noch zurückkommen. 2010 erschien außerdem das „Routledge Handbook of Corpus Linguistics". Ebenfalls aus Großbritannien kommt ein Werk, in dem Grundbegriffe (*Terms*) der Korpuslinguistik erläutert werden (Mahlberg und Brook O'Donnell, 2010). Schließlich möchten wir noch auf den sehr schönen Kurs hinweisen, den Noah Bubenhofer zusammengestellt und über das Web verfügbar gemacht hat (Bubenhofer, 2001).

Zunächst und vor allem wünschen wir Ihnen aber viel Spaß bei der Arbeit mit diesem Buch!

1.3 Aufbau des Buchs

Im zweiten Kapitel werden wir ausführlicher auf die Kritik, die von sprachtheoretischer Seite gegen die Korpuslinguistik vorgebracht wurde, eingehen. Der Gegensatz zwischen Generativer Grammatik und Korpuslinguistik ist grundsätzlich. Er wurzelt in einer unterschiedlichen Auffassung von Gegenstand und Methode der Linguistik, wie wir darstellen werden. Wir stellen die im positiven wie negativen Sinne für die Korpuslinguistik einflussreichen linguistischen Strömungen der Generativen Grammatik und des Kontextualismus vor. Am Ende dieses Kapitels werden wir drei Ansätze korpuslinguistischer Forschung gegenüberstellen: einen korpusbasierten, rein quantitativen Ansatz, einen korpusbasierten, quantitativ wie auch qualitativ ausgerichteten Ansatz und einen korpusgestützten, qualitativen Ansatz.

Im dritten Kapitel werden wir ausführlicher darstellen, was linguistische Korpora sind, in Abgrenzung zu anderen Arten linguistischer Datensammlungen. Wir werden drei für linguistische Korpora relevante Datenebenen unterscheiden: die Primärdaten, die Metadaten und die linguistische Annotation. Für die Beschreibung linguistischer Korpora haben sich auf internationaler Ebene Standards durchgesetzt. Diese Standards werden wir vorstellen. Der abschließende Teil ist methodischen Problemen gewidmet, die man lösen sollte, bevor man Korpora für eine linguistische Untersuchung heranzieht. Wir werden die folgenden Fragen beantworten: Können Korpora repräsentativ sein? Wie findet man sprachliche Phänomene in großen Mengen von Sprachdaten? Was macht man, wenn ein zu untersuchendes Phänomen nicht im Korpus gefunden wird und was, wenn man etwas findet, das auf Grund einer entwickelten Theorie eigentlich nicht vorkommen dürfte?

Linguistische Annotationen helfen, relevante Vorkommnisse in größeren Datenmengen (wieder) zu finden. Hierzu werden die Daten linguistisch voranalysiert und mit Annotationen wie zum Beispiel Wortarten oder grammatischen Funktionen versehen. Wir werden im vierten Kapitel Mittel und Methoden der Annotation darstellen und dabei unter anderem syntaktisch annotierte Korpora vorstellen. Um die Leser an die eigene Nutzung solcher Ressourcen heranzuführen, werden wir im fünften Kapitel anschließend auf die linguistische Abfrage von Korpora eingehen, darüber hinaus Methoden für das eigene Annotieren und eine Reihe von Abfrage- und Annotierwerkzeuge vorstellen.

Im sechsten Kapitel werden wir die Leser an die quantitative Auswertung auf der Basis von Korpora heranführen. Dort, wo wir grundlegende Konzepte von Mathematik und Statistik benötigen, werden wir diese informell einführen und im Übrigen auf vertiefende Literatur zu diesem Thema hinweisen. Wir, die Autoren dieses Buches, haben die Erfahrung gemacht, dass es durchaus auch Nicht-Mathematikern gelingen kann, sich das Handwerkszeug quantitativer Forschung anzueignen.

Korpora unterscheiden sich in vielfältiger Weise. Im siebten Kapitel werden wir anhand von konkreten Korpusbeispielen eine Typologie einführen, in der verschiedene Korpustypen systematisch unterschieden werden. Am Schluss dieses Kapitels stellen wir einige vielversprechende neue Korpusinitiativen vor.

Korpora sind die Materialgrundlage vielfältiger qualitativer und quantitativer sprachwissenschaftlicher Untersuchungen. Im achten und letzten Kapitel werden wir einige ausgewählte Untersuchungen präsentieren und damit die Vielfalt der Fragen sichtbar machen, die mit Hilfe von Korpora beantwortet werden können.

Glossar und Index im Anhang werden sicherlich auch denen helfen, die das Buch zum Nachschlagen oder zum Lernen auf eine Prüfung verwenden wollen. Begleitet wird dieses Buch von einer Webseite, die unter www.narr-studienbuecher.de/9783823378860 erreichbar ist. Hier finden Sie:

- Eine tabellarische Liste von Korpusprojekten. Diese Einträge werden nach den in Kapitel 7 eingeführten Kriterien beschrieben;
- Hinweise auf Werkzeuge, die die Arbeit mit Korpora erleichtern;
- Handreichungen zu einigen der gebräuchlicheren Korpuswerkzeuge;
- Lösungsansätze für die Übungsaufgaben;
- weitere nützliche Links;
- weitere Informationen zu den Autoren des Buchs.

Diese Einleitung ist ein guter Ort, um über weitere Einleitungen und Handbücher zu informieren, die unsere Leser auch interessieren könnten. Eine weitere deutsche Einführung in das Thema hat Carmen Scherer verfasst (Scherer, 2005). Dieser Text kann als eine etwas leichtgewichtigere, an Germanisten gerichtete Alternative zu diesem Buch betrachtet werden. Drei Mitarbeiter des Instituts für Deutsche Sprache (IDS) haben eine methodisch ausgerichtete Einführung erarbeitet, die den Umgang mit großen Datenmengen und die speziellen Abfragemöglichkeiten der IDS-Korpora in den Mittelpunkt stellt (Perkuhn et al., 2012). Eine in Deutsch verfasste, aber an Anglisten gerichtete Einführung ist die von Joybrato Mukherjee (Mukherjee, 2009). Korpora gesprochener Sprache, die hier nur am Rande behandelt werden, stehen im Mittelpunkt einer gut lesbaren Einführung von Christoph Draxler (Draxler, 2008). Aus dem angelsächsischen Raum ist unbedingt das Buch von Tony McEnery, Richard Xiao und Yukio Tono zu erwähnen (McEnery et al., 2006). Es ist zum einen die Fortschreibung von McEnery und Wilson (2001), zum anderen enthält sie, über das ältere Werk hinausgehend, eine Dokumentation über wichtige methodische Diskussionen innerhalb der Korpuslinguistik (Teil B) und dreizehn Fallstudien, in denen beispielhaft Schritt für Schritt korpuslinguistische Projekte entwickelt werden (Teil C). Eine weiterführende Auseinandersetzung mit den verschiedenen Strömungen der Korpuslinguistik findet sich in McEnery und Hardie (2012). An Anfänger hingegen richtet sich die Einführung von Kübler und Zinsmeister (2015), die sich auf linguistisch annotierte Korpora konzentriert.

Neben diesen Einführungen sind auch zwei Handbücher erschienen. Bei de Gruyter wurden 2008 und 2009 zwei Bände des internationalen Handbuchs „Corpus Linguistics" veröffentlicht (Lüdeling und Kytö, 2008, 2009). Auf einige Aufsätze aus diesem Handbuch werden wir im Laufe dieses Buches noch zurückkommen. 2010 erschien außerdem das „Routledge Handbook of Corpus Linguistics". Ebenfalls aus Großbritannien kommt ein Werk, in dem Grundbegriffe (*Terms*) der Korpuslinguistik erläutert werden (Mahlberg und Brook O'Donnell, 2010). Schließlich möchten wir noch auf den sehr schönen Kurs hinweisen, den Noah Bubenhofer zusammengestellt und über das Web verfügbar gemacht hat (Bubenhofer, 2001).

Zunächst und vor allem wünschen wir Ihnen aber viel Spaß bei der Arbeit mit diesem Buch!

2 Die Quellen linguistischer Erkenntnis

Nach dem Durcharbeiten dieses Kapitels werden Sie wissen, wie in zwei großen Strömungen der Linguistik, in der generativen Grammatik und im Kontextualismus, mit Sprachdaten umgegangen wurde. Sie werden die unterschiedlichen erkenntnistheoretischen Positionen, auf die beide Strömungen aufbauen, unterscheiden können und Sie werden erklären können, welches Verhältnis sie jeweils zu Sprachdaten haben und welche Arten von Sprachdaten Sie in ihrer Forschung verwenden. Sie werden verstehen, warum Noam Chomsky in einem Interview behauptete, dass es so etwas wie Korpuslinguistik nicht gebe. Sie werden aber auch gesehen haben, warum es sich dennoch lohnt, Korpuslinguistik zu betreiben. Außerdem werden Sie drei unterschiedliche Ansätze, Korpuslinguistik zu betreiben, kennengelernt haben. Sie werden Ihre eigenen Arbeiten so besser einordnen können.

Das unterschiedliche Verhältnis von Korpuslinguisten einerseits und theoretisch arbeitenden Linguisten andererseits zu Korpusdaten geht auf einen grundsätzlichen Unterschied in den erkenntnistheoretischen Grundlagen und Methoden beider Richtungen zurück. Die methodischen Grundlagen korpuslinguistischer Forschung sind *empiristisch*, die der theoretischen Linguistik *rationalistisch*. Wir wollen deshalb zunächst die erkenntnistheoretischen Grundlagen und Methoden des Empirismus und des Rationalismus darstellen, da aus der jeweiligen erkenntnistheoretischen Position ein unterschiedliches Verständnis der Rolle von authentischen Korpusdaten[1] folgt.

In den darauf folgenden Abschnitten werden wir zwei für die Korpuslinguistik bedeutende sprachtheoretische Strömungen, die generative Grammatik und den Kontextualismus, vorstellen. Es geht dabei in erster Linie um den Platz von Korpusdaten in diesen Theorien. Wir werden außerdem eine Arbeit bzw. einen Ansatz vorstellen, der dazu geeignet scheint, die Positionen dieser beiden Lager zu versöhnen und die Korpuslinguistik auf ein neues Fundament zu stellen.

Am Schluss dieses Kapitels stellen wir drei Arten, Korpusdaten für linguistische Untersuchungen zu gebrauchen, nebeneinander. Diese tabellarische Übersicht kann als Einstieg in die Fallstudien der folgenden Kapitel verwendet werden.

---

[1] Mit *authentisch* meinen wir, dass diese Daten im Rahmen linguistisch unreflektierter Kommunikationssituationen entstanden sein sollten. Es lässt sich, vor allem bei Zeitungskorpora, nicht verhindern, dass Textproduzenten sich in diesen Texten über Sprache allgemein oder einzelne sprachliche Phänomene auslassen, diese Situationen sollten allerdings eine deutliche Minderheit der ausgewerteten Belege ausmachen. Vgl. zu diesem Begriff auch Tognini-Bonelli (2001), S. 55–57.

2.1 Empirismus und Rationalismus

Es handelt sich bei Empirismus und Rationalismus um zwei erkenntnistheoretische Strömungen, deren Ursprünge bis in die antike Philosophie zurück reichen. Mit diesen Begriffen werden Ideologien bezeichnet, die vor allem in der philosophischen Debatte des 17. und 18. Jahrhunderts entschieden verfochten wurden. In der heutigen Wissenschaft spielen sie vor allem als Bedingungen der Erkenntnis eine Rolle und wirken so in den Wissenschaften, auch in der Sprachwissenschaft, weiter.

Der Kern der *empiristischen* Auffassung ist die Behauptung, dass alle Erkenntnis in der sinnlichen Anschauung wurzelt. Alles, was wir wissen können, lernen wir durch Beobachtung. Der Kern der *rationalistischen* Auffassung ist die Behauptung, dass Erkenntnisse durch Begriffe und Urteile gewonnen werden. Zu diesen gelang: man mit Hilfe der Vernunft und ohne direkten Bezug zur sinnlichen Anschauung.

Die empiristische Position lässt sich durch die folgenden Aussagen charakterisieren[2]:

- Allen *Begriffen*, die diesen Namen verdienen und die nicht bloß leere Worte sind, liegt Erfahrung zugrunde;
- *Aussagen*, die nicht aus anderen Aussagen ableitbar sind, beruhen auf Erfahrung;
- Alle Aussagen, die nicht unmittelbar auf Erfahrung beruhen, müssen aus Aussagen ableitbar sein, die dies tun.

Das erkenntnistheoretische Programm des Empirismus erfasst also sowohl Begriffe als auch Aussagen und bindet diese, direkt oder indirekt, an das, was sinnlich wahrnehmbar ist (*Erfahrung*).

Betrachten wir ein Beispiel: In der Korpuslinguistik wurde in den 90er Jahren der Begriff *Kollokation*[3] auf den Begriff der *Kookkurenz* (gemeinsames Vorkommen zweier linguistischer Einheiten, im Folgenden *Kovorkommen* genannt) zurückgeführt. Dem liegt die Einsicht zu Grunde, dass der Begriff der Kollokation nicht direkt auf Beobachtungen an Sprachdaten zurückzuführen ist. Es ist aber mittels Beobachtungen an Korpusdaten und statistischen Verfahren zu ermitteln, welche Paare von Wörtern signifikant häufiger miteinander vorkommen, als dies auf Grund einer zufälligen Verteilung von Wörtern in Texten zu erwarten wäre. Mit Hilfe dieses nun auf Beobachtungen rückführbaren Begriffs des (signifikanten) Kovorkommens wurde der Begriff *Kollokation* neu definiert. Anders ausgedrückt: Die Aussage, dass ein Wortpaar eine Kollokation ist, wird, da sie nicht direkt auf Erfahrung zurückzuführen ist, auf die Aussage gestützt, dass zwei Wörter signifikant häufig gemeinsam vorkommen, eine Aussage also, die direkt auf Erfahrung zurückführbar ist[4].

---

[2] Wir folgen hier im Wesentlichen Engfer (1996), S. 12.

[3] Beispiele für Kollokationen sind: *fieberhaft suchen, rotes Tuch, einen Antrag stellen.*

[4] Die Darstellung ist stark vereinfacht, um das Wesentliche dieses Beispiels hervorzuheben. Natürlich sind Kollokationen nicht ausschließlich durch ein quantitatives Merkmal gekennzeichnet. Wichtig ist hier, dass der Begriff *Kollokation* und Aussagen, die ihn verwenden, mittelbar auf direkte Beobachtung an Sprachdaten zurückführbar sind. Zum Verhältnis von Kollokation und Kovorkommen und zur kritischen Diskussion dieser Begriffe vor allem in der lexikographischen Literatur siehe auch Lemnitzer (1997).

Die rationalistische Position lässt sich durch die folgenden Aussagen charakterisieren[5]:

- Es wird – unter dem Titel angeborener Ideen – die Existenz erfahrungsunabhängiger Begriffe, wie *Zahl*, *Substanz* etc. angenommen;
- Es wird die Gültigkeit erfahrungsunabhängiger Aussagen behauptet. Diese beruhen allein auf vernünftiger Einsicht;
- Gestützt auf solche Aussagen oder Prinzipien lassen sich weitere Aussagen erschließen, die, wie die ursprüngliche Aussage, unabhängig von aller Erfahrung gelten.

Im rationalistischen Programm sind Begriffe und Aussagen, die sich auf Erfahrung stützen, keinesfalls ausgeschlossen. Ihnen wird aber gelegentlich gegenüber aus Vernunfteinsicht gewonnenen Begriffen und Aussagen ein geringerer Stellenwert eingeräumt.

Betrachten wir auch für diese Position ein linguistisches Beispiel: Ein in der Sprachtypologie entwickeltes Prinzip besagt, dass man Sprachen, anhand ihrer Wortstellung, unter anderem in SOV-Sprachen (Subjekt vor Objekt vor Verb) und SVO-Sprachen (Subjekt vor Verb vor Objekt) einteilen kann. Aussagen zu diesen Sprachtypen gehen auf die sprachliche Universalienforschung zurück[6]. Aus der Aussage, dass eine bestimmte natürliche Sprache eine SOV-Sprache ist, lassen sich weitere Aussagen ableiten, zum Beispiel die, dass eine auf eine Nominalphrase bezogene Präpositionalphrase der Nominalphrase folgt und ein modifizierendes Adjektiv mit hoher Wahrscheinlichkeit dem Nomen vorangeht. Das Deutsche wird von generativen Grammatikern als SOV-Sprache klassifiziert[7]. Dies deckt sich nicht unmittelbar mit Beobachtungen an deutschen Sätzen. In Beispiel (1), einem Hauptsatz, geht das Verb dem Objekt voran.

(1) Der Sprachwissenschaftler *erfindet* viele sprachliche Beispiele ...

In Beispiel (2), einem Nebensatz, folgt das Verb tatsächlich dem Subjekt und Objekt (Verbendstellung):

(2) ..., weil er Beispielen aus Korpora *misstraut*.

Aus der reinen Beobachtung und der Tatsache, dass Hauptsätze häufiger vorkommen als Nebensätze, könnte man nun schließen, dass das Deutsche tendenziell eine SVO-Sprache ist. In der generativen Grammatik wird statt dessen eine *Tiefenstruktur* angenommen, in der das Verb im Deutschen immer den Objekten folgt. In Hauptsätzen wird das finite Verb durch Transformationen oder vergleichbare Operationen an die zweite Position in der *Oberflächenstruktur* verschoben. Es spricht einiges für eine solche Argumentation. Erstens kann auch in Hauptsätzen ein Teil des Verbalkomplexes hinter den Objekten stehen:

---

5 Vgl. Engfer (1996), S. 12.

6 Vgl. Greenberg (1963). Die Universalienforschung beschäftigt sich mit den linguistischen Merkmalen, die allen Sprachen gemeinsam sind oder anhand derer sich Sprachtypen unterscheiden lassen, je nachdem, welchen Wert ein Merkmal annimmt.

7 Vgl. Grewendorf (1995): „According to the standard view, German is a ‚verb second' language whose basic (D-Structure) constituent order is verb-final.".

(3) Sie hätte den Text auch einfach gründlicher *lesen können*.

Zweitens wird die Partikel von Partikelverben dort quasi zurückgelassen:

(4) Sie hielt sich gestern mal wieder den ganzen Tag lang mit belanglosen Dingen *auf*.

Drittens ist es richtig, dass das Deutsche einige Stellungsregularitäten, zum Beispiel zwischen Adjektiv und Nomen, aufweist, die für die SOV-Sprachen charakteristisch sind.

Die Aussagen zu SOV- und SVO-Sprachen sind somit nicht auf Erfahrung zurückführbar, denn Tiefenstrukturen sind der unmittelbaren Beobachtung nicht zugänglich. Auch Begriffe wie *Subjekt* und *Objekt* sind keine Erfahrungsbegriffe. Sie sind das Ergebnis vernunftgeleiteter Überlegungen. Die Stärke der verwendeten Begriffe und Aussagen liegt darin, dass sie Zusammenhänge zwischen Phänomenen erklären können.

Im Allgemeinen wird der Empirismus als Erkenntnistheorie mit der wissenschaftlichen Methode der *Induktion* und der Rationalismus mit der wissenschaftlichen Methode der *Deduktion* verbunden. Die Induktion lässt sich als Schlussverfahren wie folgt charakterisieren:

- Übergang vom Besonderen zum Allgemeinen;
- Schließen von einzelnen Beobachtungen auf Gesetzesaussagen;
- Möglichkeit der Widerlegung von Gesetzesaussagen durch Beobachtungen.

Die Deduktion lässt sich wie folgt charakterisieren:

- Übergang vom Allgemeinen zum Besonderen;
- Schluss von Prinzipien und Axiomen auf Regeln;
- Möglichkeit der Überprüfung der Gültigkeit dieser Regeln durch Beobachtungen.

Auch dies möchten wir an einem linguistischen Beispiel veranschaulichen: Aus der Beobachtung, dass *einige finite Verbformen Bestandteile von Hauptsätzen sind*, und der Beobachtung, dass *diese finiten Verbformen an zweiter Stelle im Satz stehen*, wird durch Induktion die Gesetzesaussage abgeleitet, dass *finite Verben in Hauptsätzen immer an zweiter Stelle stehen*. Diese kann an Beobachtungen überprüft und falsifiziert[8] werden. So trifft die Aussage z.B. für den Satz in Beispiel (5) nicht zu:

(5) Bleib wo du bist!

Auf Grund dieser und weiterer, der Gesetzesaussage widersprechender Evidenz kann diese verworfen oder modifiziert werden. Die Aussage kann z.B. eingeschränkt werden: *finite Verben in den Hauptsätzen, die Aussagesätze sind, stehen immer an zweiter Stelle*.

Anders herum kann aus dem unabhängig motivierten Prinzip der SOV- und SVO-Stellung von Konstituenten in Sätzen und der Feststellung, dass das Deutsche eine SOV-Sprache ist, deduktiv geschlossen werden, dass das finite Verb am Satzende steht. Die

[8] Mit *Falsifikation* wird das Verfahren bezeichnet, eine Gesetzesaussage durch mindestens ein Gegenbeispiel zu widerlegen bzw. zu verwerfen. In statistischer Ausdrucksweise würde hierfür eine signifikante Anzahl von Gegenbeispielen benötigt.

beobachtbare Tatsache, dass im Deutschen in Aussagesätzen das Verb an zweiter Stelle steht, wird mit der Regel dadurch in Einklang gebracht, dass eine Transformation angenommen wird, die das finite Verb aus der Endstellung in einer Tiefenstruktur an die zweite Position in der Oberflächenstruktur bewegt.

Im Rahmen rationalistisch orientierter sprachwissenschaftlicher Forschung kann ein Korpus zur Überprüfung und Korrektur theoretischer Aussagen verwendet werden. Wir werden dies *korpusgestützte* Linguistik nennen, da das Korpus hier primär für die Stützung von im Vorhinein entwickelten Hypothesen herangezogen wird.

Im Rahmen empiristisch orientierter sprachwissenschaftlicher Forschung ist das Korpus die primäre Quelle der Erkenntnis. Aus Beobachtungen an authentischen Sprachdaten werden Gesetzesaussagen abgeleitet, die durch weitere Beobachtungen bestätigt, modifiziert oder verworfen werden. Wir werden dies *korpusbasierte* Linguistik nennen, da das Korpus als die Basis der Erkenntnis, also auch der Bildung von Theorien und Hypothesen, herangezogen wird[9].

2.2 Sprecherurteile statt Korpusdaten — Die Position der generativen Grammatik

Alle sprachwissenschaftliche Forschung bezieht sich auf sprachliche Daten. Nur als eine Menge von gesprochenen oder geschriebenen Äußerungen kann sich das Sprachvermögen als kognitive Leistung von Menschen oder das System einer natürlichen Sprache manifestieren. Schon Leonard Bloomfield stellte in den zwanziger Jahren des letzten Jahrhunderts in einem programmatischen Aufsatz fest, dass die Gesamtheit der Äußerungen, die in einer Sprachgemeinschaft gemacht werden können, die Sprache dieser Sprachgemeinschaft sei[10].

Bei dieser und bei ähnlichen Formulierungen zur Gegenstandsbestimmung der Sprachwissenschaft setzt nun die Kritik der generativen Grammatik[11] an, die seit den fünfziger Jahren das Forschungsprogramm der Sprachwissenschaft prägt. Die *Gesamtheit der Äußerungen* sei eine fiktive Größe, die im Fall einer lebenden, aktuell verwendeten Sprache durch keine Kollektion von Äußerungen auch nur annähernd repräsentiert werden könne. Eine Sprache durch Aufzählung aller Äußerungen erfassen zu wollen, sei nicht nur ein äußerst langweiliges, sondern auch ein müßiges Unterfangen. An dieser Stelle wird oft eine Analogie zum Schachspiel bemüht: Man lernt und versteht dieses Spiel nicht, wenn man die Zugfolgen möglichst vieler Partien betrachtet, sondern nur, indem man einige wenige Regeln lernt und diese anwendet. In ähnlicher Weise wird in der generativen Grammatik als eigentlicher Gegenstand der Forschung die kognitive

---

[9] Diese Unterscheidung geht im Wesentlichen auf Elena Tognini-Bonelli (2001) zurück. Diese verwendet den Ausdruck ‚corpus-driven' für den Ansatz, den wir hier korpusbasiert nennen und der an anderer Stelle auch „korpusgeleitet" genannt wird. Für das, was wir hier „korpusgestützt" nennen, verwendet sie den Ausdruck ‚corpus-based'. Die Leser sollten sich hier nicht verwirren lassen.

[10] Vgl. Bloomfield (1926), S. 153.

[11] Als *generative Grammatik* wird ein Grammatikmodell bezeichnet, nach dem durch ein begrenztes Inventar von Regeln alle wohlgeformten Sätze einer Sprache generiert werden können. Der Begriff *generative Grammatik* bezeichnet außerdem eine sprachwissenschaftliche Schule, in der dieses Grammatikmodell eine zentrale Rolle spielt.

Maschinerie (‚generative device') angesehen, die es Menschen ermöglicht, mit einem begrenzten Inventar von Regeln eine theoretisch unbegrenzte Menge von Äußerungen zu produzieren. Die Gesamtheit der bereits irgendwann getätigten Äußerungen sei für die Beschreibung bzw. Erklärung dieser kognitiven Maschinerie irrelevant.

Chomsky hat zwei Begriffspaare für die Dichotomie von konkreten sprachlichen Äußerungen einerseits, und der Fähigkeit sich sprachlich zu äußern andererseits, verwendet: zunächst *Performanz* und *Kompetenz*, später *E-Sprache* und *I-Sprache*.

Wir werden im Folgenden kurz die Dichotomie von Kompetenz und Performanz einführen und dann ausführlicher auf die Argumentation Chomskys eingehen, mit der er den Unterschied von E-Sprache und I-Sprache begründet[12].

Betrachten wir zunächst das Begriffspaar *Kompetenz* und *Performanz*[13].

Definition 1 (Performanz). *Performanz, die auch* Sprachverwendung *genannt wird, ist der aktuelle Gebrauch der Sprache in konkreten Situationen.*

Definition 2 (Kompetenz). *Die Kompetenz eines (idealen) Sprechers ist das ihm angeborene oder von ihm erworbene sprachliche Wissen. Dieses umfasst ein System von Prinzipien und Regeln. Dieses Wissen ermöglicht es dem Sprecher, eine im Prinzip unendliche Menge von Äußerungen hervorzubringen und zu verstehen, Urteile über die Wohlgeformtheit von Äußerungen zu treffen sowie die Mehrdeutigkeit oder die Bedeutungsgleichheit von Sätzen zu erkennen.*

Die Kompetenz von Sprechern ist ein theoretisches Konstrukt, etwas, zu dem Forscher keinen unmittelbaren Zugang haben. Die Performanz hingegen ist als Menge von Äußerungsereignissen der Beobachtung unmittelbar zugänglich. Sprachwissenschaftler, die im theoretischen Rahmen der generativen Grammatik arbeiten, bestreiten, dass sich aus der beobachteten Sprachverwendung Schlüsse auf die Kompetenz ziehen lassen. Die sprachliche Leistung von Sprechern, ihre Performanz, wird durch vielfältige Faktoren beeinflusst, die nichts mit dem Sprachvermögen zu tun haben, zum Beispiel durch Begrenzungen des Kurzzeitgedächtnisses, momentane Unaufmerksamkeit und äußere Ablenkungen.

So würde der tatsächlich belegte Satz:

(6) Anstelle des alten Magazins entstand vor einem Jahr ein fensterloser Trumm, *in dem* erst das Großkino „CinemaxX" einzog und nun auch das „Übermaxx" residiert ... (taz, 30. April 1999)

von den meisten deutschen Muttersprachlern, wenn er ihnen vorgelegt werden würde, als ungrammatisch empfunden – das Verb *einziehen* verlangt eine Präpositionalphrase mit einer Nominalphrase (NP) im Akkusativ als Komplement, nicht, wie im obigen Beispiel, einer NP im Dativ. Eine grammatische Beschreibung des Verbs *einziehen* würde aber, wenn sie sich auf diesen Beleg stützte, eine Präpositionalphrase mit einer NP im

---

[12] Eine gründliche Analyse des Korpusbezugs in Chomskys früheren Arbeiten, bis zu denen der späten sechziger Jahre, hat Fred Karlsson (2008) vorgelegt. Seine Schlussfolgerungen entsprechen weitgehend den hier vorgestellten.

[13] Wir beziehen uns im Folgenden auf Chomsky (1969), Kapitel 1, § 1.

Dativ als Komplement zulassen. Man könnte einwenden, dass die Beschreibung sprachlicher Phänomene sich nicht auf eine einzelne Beobachtung stützen sollte. Die Vorkommenshäufigkeit eines Phänomens spielt also eine wichtige Rolle.

Der Fehler im folgenden Beleg ist vermutlich kein Einzelfall:

(7) Allerdings haben die Bremer am 11. Mai noch ein Nachholheimspiel gegen Schalke 04, *daß* aus Sicherheitsgründen abgesagt wurde. (taz, 4.5. 1999)

Das Relativpronomen *das* und die subordinierende Konjunktion *daß* (*dass* in neuer Rechtschreibung) werden häufig verwechselt, in beide Richtungen. Aus Belegen wie in Beispiel (7) darf nun nicht der Schluss gezogen werden, dass das Lexem *dass* als Relativpronomen verwendet werden kann. Für unser Wissen als Muttersprachler des Deutschen stellt dies kein Problem dar, wohl aber für eine Sprachbeschreibung, die sich ausschließlich auf die Produkte der Performanz stützt. Beispiele wie diese begründen die Skepsis vieler Sprachwissenschaftler gegenüber authentischen Sprachdaten als Schlüssel zur Erkenntnis des sprachlichen Wissens. Eine performanzorientierte Sprachwissenschaft muss deshalb die folgenden Fragen beantworten können: Ist eine Konstruktion grammatisch, obwohl sie nur selten vorkommt? Welche Konstruktionen sind ungrammatisch, obwohl sie häufig verwendet werden?

Performanzdaten helfen, so die generativen Grammatiker, bei der Bestimmung der Sprachkompetenz nicht weiter, da sie durch die genannten Faktoren „verunreinigt" sein können. Nur Sprecherurteile, also Selbstauskünfte von Sprechern über ihr sprachliches Wissen, sind in diesem theoretischen Rahmen als Primärdaten zugelassen. Es könnte zum Beispiel Gegenstand der Untersuchung sein, herauszufinden, welche Sätze Sprecher des Deutschen als ungrammatisch charakterisieren würden.

(8) *Peter wohnt.

(9) ?Peter wohnt mal wieder.

(10) Peter wohnt komfortabel.

(11) Peter wohnt in Berlin.

Das durch den Stern und das Fragezeichen markierte Sprecherurteil ist für diese Zwecke erfunden, aber sicher leicht nachvollziehbar. Offenbar verlangt das Verb *wohnen* nach einem modalen oder lokalen Adverb als Ergänzung (Beispiele (10) und (11)). Ein iteratives Adverb ist schon deutlich fragwürdiger (Beispiel (9), das deshalb mit einem Fragezeichen gekennzeichnet ist). Ohne weitere Ergänzung außer dem Subjekt ist der Satz aber ungrammatisch (Beispiel (8), der Stern markiert den Verstoß des Beispiels gegen grammatische Regeln).

In späteren Arbeiten führt Chomsky eine weitere Unterscheidung ein, die zwischen E-Sprache und I-Sprache. Wir beziehen uns im Folgenden auf Chomskys Essay *Knowledge of Language. Its Nature, Origin, and Use*[14]. Chomsky charakterisiert sein Forschungsprogramm als Abstraktion weg vom konkreten sprachlichen Verhalten bzw. von dessen

[14] Vgl. Chomsky (1986). Die Zahlen in Klammern geben die Seitenzahlen an, auf die wir uns beziehen.

Produkten und hin zu den mentalen Zuständen, die dieses Verhalten bestimmen. Die Aufgabe der Sprachwissenschaft ist es, Antworten auf die folgenden Fragen zu finden:

1. Woraus besteht unser Sprachwissen?
2. Wie wird es erworben?
3. Wie wird es angewendet? (3)

Chomsky kritisiert explizit die beschreibende und strukturalistische Sprachwissenschaft und die Verhaltenspsychologie dafür, dass sie Sprache als eine Reihe von Sprachhandlungen oder als eine Menge sprachlicher Formen, gepaart mit Bedeutungen betrachtet haben (19). Diese Kritik trifft sicher auf die Form von empirischer Sprachwissenschaft zu, wie sie Bloomfield in dem oben dargestellten Sinn skizzierte. Die Menge der Äußerungsereignisse oder Sprachhandlungen bezeichnet Chomsky als *E-Sprache* ('E-language', 20), als externalisierte Sprache in dem Sinne, dass sie nicht in Zusammenhang mit mentalen Zuständen der Sprecher betrachtet wird. Eine Grammatik, die aus diesen Daten abgeleitet werden würde, stelle nicht mehr als eine Sammlung von Beschreibungen dieser Ereignisse und Handlungen dar. Eine solche Grammatik wäre ein arbiträres Gebilde, deren einziges Qualitätskriterium es ist, die beobachteten sprachlichen Ereignisse korrekt zu beschreiben (20).

Dem stellt Chomsky die *I-Sprache* ('I-language', 22) gegenüber. Mit dem Ausdruck *internalisierte Sprache* bezeichnet Chomsky mentale Zustände der Sprecher, die eine Sprache beherrschen (22). Eine Grammatik ist eine Theorie über diese I-Sprache und damit über die mentalen Zustände der Sprecher. Grammatiken, verstanden als Theorien über die I-Sprache, sollen so einfach wie möglich sein. Sie sind außerdem falsifizierbar wie jede andere wissenschaftliche Theorie. Dies sind die wissenschaftlichen Kriterien, nach denen Grammatiken bewertet werden können, wenn mehrere gleichermaßen die I-Sprache beschreiben. Die Konstruktion und Auswahl einer Theorie ist also keinesfalls willkürlich.

Für den Erkenntniswert der Korpuslinguistik bedeutet dies: Selbst wenn man, auf Grund eines ausreichend großen Korpus, zuverlässige Aussagen über die möglichen Ausdrücke einer natürlichen Sprache, also über die E-Sprache, erlangen könnte, wäre dies nicht ausreichend für die Bestimmung der I-Sprache, da es mehr als eine interne Sprache geben könnte, die exakt dieselben möglichen Ausdrücke erzeugt. Die konkret beobachtbaren Äußerungen liefern außerdem keinen Schlüssel zu den mentalen Zuständen der Sprecher, die nach Chomsky der eigentliche (und ausschließliche) Gegenstand der Sprachwissenschaft sein sollen.

Wie ist nun aber der Zugang zu den mentalen Zuständen der Sprecher möglich? Chomsky schlägt folgende Quellen vor:

- Die wichtigste Quelle ist das Sprachgefühl bzw. Intuition (engl: ‚intuition') der Sprecher, die direkt oder indirekt über ihr Sprachwissen Auskunft geben[15]. Sprecher können direkt Auskunft geben, indem sie z.B. die Grammatikalität oder Akzeptabilität von Sätzen beurteilen, die ihnen vorgelegt werden, oder angeben, ob sie selber

[15] „Hill: If I took some of your statements literally, I would say that you are not studying language at all, but some form of psychology, the intuitions of native speakers. Chomsky: That is studying language.", zit. nach Harris (1995), S. 54.

einen solchen Satz verwenden würden. Indirekte Auskunft kann dadurch eingeholt werden, dass Sprecher in Experimente einbezogen werden, in deren Verlauf ihnen bestimmte Äußerungen entlockt (engl. ‚elicit') werden[16].

- Darüber hinaus können die folgenden Quellen indirekt zu Erkenntnissen über das Sprachvermögen beitragen[17]:
  - Befunde über Sprachstörungen (Stottern, Aphasien usw.);
  - Versprecher (*Sehr geehrte Damen und Herren*)[18];
  - Neu geprägte Sprachen, z.B. die Kreolsprachen[19].

  Sprachstörungen sind ein indirekter Beleg für den modularen Aufbau des Sprachvermögens, denn bei den meisten Sprachstörungen sind nur einige Bereiche oder Aspekte des Sprechens gestört bzw. des Schreibens, wie im Falle der Legasthenie. Anhand von neurologischen Befunden, zum Beispiel Hirnläsionen nach einem Unfall, die mit dem Ausfall bestimmter sprachlicher Fähigkeiten korrespondieren, lässt sich der Sitz des Sprachvermögens im Gehirn nachweisen. Versprecher deuten auf momentane Fehlfunktionen auf dem Wege von der Planung zur Realisierung einer Äußerung hin. Die Art der Fehlfunktion erlaubt wiederum Rückschlüsse auf den modularen Charakter des Sprachvermögens. Es kann gezeigt werden, dass bei Versprechern bestimmte Aspekte der Sprachproduktion in systematischer Weise gestört werden[20]. Kreolsprachen als eine Verfestigung des Vermischungsprozesses mehrerer Sprachen, unter deren Einfluss die Sprecher standen (z.B. indigene Sprache und Amtssprache), lassen prinzipiell Schlüsse auf die Erlernbarkeit von Sprachen zu.

Gegenüber diesen Quellen linguistischer Erkenntnis leiden Korpora unter den folgenden Mängeln:

- Korpora enthalten eine nicht unerhebliche Anzahl von Äußerungen, die von Sprechern, wenn sie diese Äußerungen zu beurteilen hätten, als nicht wohlgeformt eingestuft würden. Ursache für diese Einstufungen können banale Dinge wie Kongruenzfehler oder Wortauslassungen sein. Es kann sich aber auch um sehr subtile Phänomene handeln, deren (Nicht-)Wohlgeformtheit nicht einfach und einhellig festgestellt werden kann. Es sind diese subtilen (Pseudo-)Fehler, die die Interpretation von Korpusdaten besonders erschweren. Eine Grammatik einer Sprache, die sich ausschließlich, ohne ein weiteres Korrektiv, auf Korpusdaten dieser Sprache stützen würde, müsste solche Sätze wie in Beispiel (7) aufnehmen und grammatisch beschreiben[21].

---

16 Labov (1975) stellt einige dieser Experimente vor, z.B. Seite 18ff. und Seite 49ff.

17 Vgl. Chomsky (1986), S. 37.

18 Vgl. Bierwisch (1970) und Leuninger (1996).

19 Kreolsprachen sind Mischsprachen in Zonen intensiven Austauschs zwischen zwei Sprachgemeinschaften. Im Gegensatz zum *Pidgin* haben diese Sprachen bereits den Charakter von Muttersprachen, d.h. es gibt bereits Sprecher, die mit dieser Sprache aufgewachsen sind; zu den Pidgin- und Kreolsprachen vgl. Camp und Hancock (1974) und Bickerton (1984).

20 Für eine detaillierte Analyse vgl. Leuninger (1996).

21 Ein weiteres, berühmtes Beispiel ist der Satz *Ich habe fertig*, im Jahr 1998 geäußert vom italienischen Trainer Giovanni Trapattoni, ehemals Trainer des FC Bayern München, auf einer Pressekonferenz. Die Ursachen für diesen Fehler liegt in der mangelnden Beherrschung der

- Selbst in den größten Korpora wird man eine Menge sprachlicher Phänomene, die für den Entwurf einer Grammatik der zu beschreibenden Sprache wichtig sind, nicht finden. Dies ist seit dem Aufkommen der generativen Grammatik eine Binsenweisheit. In jedem neuen Text wird man Sätze finden, die vorher noch nie geäußert bzw. aufgeschrieben wurden. Was für einzelne Sätze gilt, kann aber auch auf Konstruktionstypen zutreffen, und dieser Mangel ist für die Beschreibung von Sprachen oder gar für die Theoriebildung viel gravierender. Wenn in einem Korpus der deutschen Sprache keine Imperativform (wie z.B. *Gib!*) oder keine sogenannte Mittelkonstruktion (wie z.B. *Dieses Auto fährt sich gut*) auftauchte, dann könnten diese Konstruktionstypen auch nicht in einer rein empirischen, korpusbasierten Grammatik erfasst werden. Sprecher des Deutschen würden diese Konstruktionen, wenn man sie ihnen vorlegte, aber sicher als wohlgeformt einstufen und sie bei gegebenem Anlass auch selber verwenden. Ihr Fehlen im Korpus ist ein reines Zufallsprodukt.

Generative Grammatiker bemühen lieber ihr eigenes Sprachgefühl, um über die Möglichkeit oder Wohlgeformtheit bestimmter Konstrukte in einer Sprache zu urteilen. So behauptete Chomsky selber in einer Diskussion, dass das Verb *perform* nicht mit unzählbaren Substantiven ('mass nouns') verwendet werden kann (**perform labour*), sondern nur mit zählbaren Substantiven ('count nouns' – *perform a task*). Er beruft sich darauf, dass er dies als Muttersprachler des Englischen wisse. Tatsächlich ist diese Verallgemeinerung falsch. Ein Blick in das *British National Corpus* zeigt (als Gegenbeispiel) die Konstruktion (*to*) *perform magic*[22].

Labov[23] führt einen extremeren Fall eines irregeleiteten Sprecherurteils an. Sprecher des amerikanischen Englisch aus Philadelphia wurden zum (korrekten) Gebrauch des Wortes *anymore* befragt. Wurden die Sätze mit diesem Wort vorgelegt, gaben viele von ihnen an, dass sie das Wort so wie in Beispiel (12) verwendet noch nie gehört hätten und dass sie dies nicht als korrektes Englisch akzeptieren könnten.

(12) John is smoking a lot anymore.

Einige interpretierten auch die Bedeutung dieses und ähnlicher Sätze falsch. Alle Kriterien deuteten also darauf hin, dass diese Konstruktionen nicht zur Sprachkompetenz dieser Sprecher gehören. Tatsächlich aber wurden diese Probanden beobachtet, wie sie dieses Wort in ähnlichen Konstruktionen verwendeten, zum Teil sogar in denselben Interviews, in denen sie zur Verwendung dieses Wortes befragt wurden[24].

Es gibt in der Literatur noch mehr Beispiele, die die Unzuverlässigkeit von Sprecherurteilen eindrücklich belegen[25]. Auch wenn Sprachwissenschaftler als Fachleute, die den reflektierten Umgang mit Sprache ihr ganzes berufliches Leben über trainieren,

---

deutschen Sprache. Der Satz hat aber durch häufige Pressezitate mittlerweile den Status eines geflügelten Wortes. Man wird ihm in Zeitungstexten dieser Zeit sicher häufig begegnen. Aber will man diesen Satz wirklich als *wohlgeformt* akzeptieren und beschreiben?

22 Das Beispiel stammt aus McEnery und Wilson (1996), S. 11.

23 Vgl. Labov (1975), S. 34f.

24 Ähnlich könnte es deutschen Sprechern, die aus dem Ruhrgebiet stammen, mit dem Satz *Ich war meine Reise am Planen* ergehen.

25 Einige wichtige Studien werden in Labov (1975) diskutiert.

wohl als die besseren Informanten gelten können, sind auch sie nicht vor Fehlurteilen sicher, wie das obige Beispiel zeigt[26].

Chomsky selber schätzt denn auch den Wert von Sprecherurteilen als linguistische Daten kritisch ein. Er möchte den grundlegenden Aufbau der Grammatik einer Sprache auf die eindeutig entscheidbaren Fälle stützen. Ist erst einmal eine solche Grundgrammatik gefunden, die die eindeutigen Fälle von wohlgeformten Sätzen einschließt und die eindeutig nicht-wohlgeformten Sätze ausschließt, dann könne aus dieser Grammatik auch der Status – wohlgeformt oder nicht – der zweifelhaften Konstruktionen abgeleitet werden[27]. Außerdem bedürften auch Sprecherurteile der Interpretation, da sie nicht direkt die Struktur der untersuchten Sprache und ihre Grammatik reflektierten[28].

Die methodische Vorsicht gegenüber Sprecherurteilen ist, wie wir gesehen haben, sicher angebracht. Es ist aber zweierlei gegen Chomskys Vorgehen vorzubringen. Erstens kann man fragen, warum man in den eindeutigen Fällen nicht auch auf Korpusdaten zurückgreifen können sollte. Die eindeutigen Fälle dürften auch die sein, die in einem großen Korpus so oft vorkommen, dass die Gefahr der Missinterpretation von Performanzfehlern gering ist. Zweitens ist der sprachtheoretische Diskurs über die korrekte Grammatik einer Sprache mittlerweile so komplex, dass vor allem über seltene Konstruktionen und deren grammatischen Status diskutiert wird. Eine konkrete Grammatik muss sich gerade an diesen Beispielen beweisen[29]. Für diese Satztypen ist aber nicht nur die Evidenz in Korpora rar und möglicherweise fragwürdig, auch das Sprachgefühl wird hier unscharf und die geforderte Konsistenz von Sprecherurteilen schwindet.

Wir möchten allerdings darauf hinweisen, dass sich auch die wissenschaftliche Praxis der Ermittlung von Sprecherurteilen verbessert hat. Dies ist ein durchaus spannendes Feld linguistischer Forschung, welches allerdings außerhalb des Rahmens dieses Buches liegt[30].

Diese kritische Bewertung introspektiver Daten als Quelle linguistischer Erkenntnis soll nicht davon ablenken, dass auch Korpusdaten problematisch sein können. Die Kritik an dem Wert von Korpusdaten sei hier noch einmal in vier Punkten zusammengefasst:

1. Der Status eines beliebig großen Korpus zu der Sprache, die es repräsentieren soll, ist unklar, da die repräsentierte Sprache aus einer potenziell unendlichen Menge von Sätzen besteht (Problem der Repräsentativität);
2. Ein Korpus enthält eine große Zahl von Phänomenen, die für die Beschreibung der Sprache, die es repräsentiert, irrelevant sind (Problem der Relevanz der Daten);

---

[26] Ein extremer Fall linguistischer Fehleinschätzung, betreffend die Möglichkeit der Einbettung von Konstituenten in Konstituenten des gleichen Typs ('central embedding'), bewog Geoffrey Sampson einst dazu, in das Lager der Korpuslinguistik zu wechseln, vgl. Sampson (1996).

[27] Vgl. Chomsky (1957), S. 14: „In many intermediate cases we shall be prepared to let the grammar itself decide".

[28] Vgl. Chomsky (1986), S. 36.

[29] Vgl. Labov (1975), S. 17: „... the acceptability of complex sentence types frequently becomes a turning point for a theoretical conclusion."

[30] Wir empfehlen dem interessierten Leser einige neuere und interessante Arbeiten zu diesem Thema: Featherston (2007), Featherston (2009) und Meyer (2009).

3. Viele Konstruktionen, die im Beschreibungsbereich einer Grammatik liegen, da sie wohlgeformt sind, sind in Korpora dieser Sprache nicht vorhanden (Problem unvollständiger Datenabdeckung);
4. Viele Äußerungen, die dann auch Bestandteile von Korpora sein können, sind nicht wohlgeformt. Aus ihnen können und sollten keine Schlüsse auf das sprachliche Wissen der Sprecher gezogen werden (Problem der Verlässlichkeit der Daten).

Man sollte als Sprachwissenschaftler, der mit Korpora arbeitet, diese Kritik an Korpusdaten ernst nehmen und dieser Kritik mit guten Argumenten begegnen können. Hierzu gehören Antworten auf die Fragen, wie man mit der Existenz nicht-wohlgeformter Äußerungen und mit dem Fehlen wohlgeformter Äußerungen umgeht. Zweifelhafte Schlüsse können zum Beispiel durch andere Daten, wie Sprecherbefragungen, gestützt werden.

In einer neueren Arbeit zu den wissenschaftstheoretischen Grundlagen der Korpuslinguistik, wird das Konzept der *Plausibilität* einer auf Korpusevidenz basierenden Aussage eingeführt[31]. Ausgangspunkt ist die Überlegung, dass angesichts der auch in diesem Kapitel dargestellten Problematik aller Korpusevidenz es nicht möglich ist, eine theoretische Aussage, die auf diesen Daten fußt, zu falsifizieren. Die Daten, die die Grundlage einer Falsifizierung bilden könnten, sind letztlich genauso problematisch, wie die Daten, auf denen die ursprüngliche Aussage beruht. Die Autoren führen deshalb einen Formalismus ein (vor allem in Kapitel 9 und 10), mit dessen Hilfe die Plausibilität einer sich auf Korpusevidenz stützenden Theorie oder Aussage quantifiziert werden kann. In die Berechnung geht auch die Evidenz ein, die gegen eine linguistische Aussage oder Theorie ins Feld geführt wird. Diese Gegenevidenz, genauer: die Evidenz, die für eine konkurrierende Theorie oder Hypothese ins Feld geführt wird, kann die Plausibilität der ursprünglichen Hypothese verringern und zu Revisionen der linguistischen Aussage oder Theorie führen. Im Extremfall muss die ursprüngliche Theorie verworfen werden[32]. Die zentralen Aussagen einer Metatheorie der plausiblen Argumentation in der Korpuslinguistik sind die Folgenden: a) der Pluralismus verschiedener Datentypen (Korpusdaten, Sprecherurteile u.a.) wird anerkannt; es sollte versucht werden, verschiedene Quellen linguistischer Evidenz heranzuziehen, um eine linguistische Aussage oder Theorie zu stärken; b) alle Quellen linguistischer Evidenz werden a priori als problematisch angesehen. Über den Wert jeder einzelnen Quelle für die jeweilige linguistische Aussage ist Rechenschaft abzulegen; c) das Verhältnis zwischen Daten und Theorie ist zyklisch. Neue Daten können zu Modifikationen der Theorie führen und eine modifizierte Theorie einen neuen Blick auf die Daten eröffnen; d) die für oder gegen eine Theorie herangezogenen Daten sind immer vorläufig und können in ihrer Gänze sogar widersprüchlich sein. Diese Widersprüchlichkeit muss eine plausible linguistische Aussage oder Theorie berücksichtigen[33].

Diese Version eines erklärenden Ansatzes trägt anders als bisherige Ansätze dem Umstand Rechnung, dass Korpusdaten den von ihnen beschriebenen Gegenstand nicht

---

31 Im Weiteren folgen wir Kertész und Rákosi (2012), vor allem Kapitel 6, S. 41ff.

32 Die Autoren stellen das ausführlich an einem Beispiel aus der Phonologie des Deutschen dar, vgl. Kertész und Rákosi (2012), S. 178–183.

33 Vgl. Kertész und Rákosi (2012), S. 41.

repräsentativ abbilden können und lückenhaft und sogar widersprüchlich sein können. Dennoch ist es nach diesem Ansatz möglich, linguistische Theorien auf ihnen zu errichten, nötigenfalls umzubauen und so zum Erkenntnisfortschritt in den Sprachwissenschaften beizutragen. Ob sich das Konzept der Plausibilität wirklich für eine Quantifizierung eignet, einer Theorie oder Aussage sich also ein Wert auf einer Plausibilitätsskala zuordnen lässt, das muss sich erst noch in weiteren Studien und anhand weiterer Beispiele erweisen. Es reicht womöglich, einer Aussage oder Theorie ein eher vage quantifizierendes Prädikat ('sehr' oder 'wenig' plausibel) oder ein komparatives Prädikat ('A ist (angesichts der Evidenz) plausibler als B') zuzuschreiben.

2.3 Linguistische Erkenntnis geht vom Sprachgebrauch aus — Die Position des Kontextualismus

Dieser Abschnitt ist einer linguistischen Schule gewidmet, für die die Arbeit mit Korpusdaten notwendiger Bestandteil linguistischer Erkenntnis ist. Für diese Schule, die in Deutschland *Kontextualismus* genannt wird[34], geht alle linguistische Erkenntnis vom Sprachgebrauch aus.

Einige prominente Korpuslinguisten, zum Beispiel John Sinclair – ehemaliger Chefredakteur des *Collins Cobuild English Dictionary* – entstammen der Schule des Kontextualismus. Für diese Sprachwissenschaftler ist die Arbeit mit sehr großen Textkorpora der Vollzug des Forschungsprogramms, das vor allem von John Rupert Firth entworfen wurde[35].

Das Forschungsziel des Kontextualismus ist es, sprachliche Äußerungen und deren verschiedenen linguistischen Aspekte als Funktionen des sprachlichen und nichtsprachlichen Kontextes zu erklären, in dem diese Äußerungen stehen.

Der erste prinzipielle Unterschied zwischen Kontextualismus und generativer Grammatik liegt in der Bestimmung des Untersuchungsgegenstandes: Während es letzterer um die Kompetenz von Sprechern und damit um die Voraussetzungen für die Bildung sprachlicher Ausdrücke geht, untersucht der Kontextualismus konkrete Verwendungsweisen von Sprache anhand von tatsächlich vorkommenden Äußerungen. Nur für konkrete Äußerungen lässt sich ein Kontext ermitteln und somit das Verhältnis zwischen Äußerung und Kontext. Experimentelles Vorgehen, z.B. Transformations- und Ersetzungstests, und die Erhebung introspektiver Daten werden abgelehnt.

Auch die Kontextualisten möchten letztendlich zu Aussagen über das Sprachsystem gelangen. Soweit scheinen der Kontextualismus und die generative Grammatik übereinzustimmen. Ein wesentlicher Unterschied liegt allerdings im Verständnis dessen, was als *Sprachsystem* bezeichnet wird. Für die generative Grammatik ist dies eine kognitive Struktur, das sprachliche Wissen der Sprecher. Für den Kontextualismus sind dies die regelhaften Beziehungen zwischen der Form, dem Inhalt und dem Kontext sprachlicher Äußerungen. Diese Beziehungen können nur aus konkreten Sprachhandlungen abstra-

---

[34] Im englischen Sprachraum sind die Bezeichnungen *London School* – der Hauptvertreter lehrte in London – oder *Functionalism* gebräuchlicher, vgl. Lehr (1996), S. 7.

[35] Wir stützen uns bei der folgenden Darstellung vor allem auf die vorzügliche Darstellung des Kontextualismus bei Andrea Lehr (1996), sowie auf die Arbeiten von Elena Tognini-Bonelli (2001).

hiert werden. Am Beginn dieser Abstraktion muss also die Erfassung, Analyse und Systematisierung der konkreten Sprachhandlungen stehen. Für Firth liegt die Bedeutung linguistischer Einheiten in ihrer Funktion für den *Kontext*, in den die Äußerungen eingebettet sind.

Definition 1 (Kontext). *Der Kontext einer Äußerung ist die Summe der unmittelbaren Rahmenbedingungen einer Sprachhandlung als das Bezugssystem, innerhalb dessen einer Äußerung eine Funktion zukommt. Dabei bildet der* kulturelle Kontext *das Bezugssystem für eine Sprache und steuert die Art und Weise, wie Sprecher sprachliche Handlungen wahrnehmen. Der* situative Kontext *determiniert die Funktion einer konkreten sprachlichen Handlung. Zum situativen Kontext gehören Ort und Zeit, die Beteiligten etc.*[36]

Das Konzept des Kontexts als Rahmen und Bedingung menschlichen Handelns hat Firth von dem Anthropologen Bronislaw Malinowski übernommen. Er hat dieses System auf linguistische Untersuchungen hin erweitert, indem er dem – im Wesentlichen nichtsprachlichen – Kontext das Konzept des Kotextes an die Seite stellte[37].

Definition 2 (Kotext). *Der Kotext einer linguistischen Einheit ist die Menge der linguistischen Einheiten, die im gleichen Text verwendet wurden. Diese linguistischen Einheiten determinieren die Funktion und die Bedeutung der untersuchten Einheit.*

Ko- und Kontext spielen für die Untersuchung sprachlicher Handlungen eine zentrale Rolle. Sie haben die deutsche Bezeichnung für diese linguistische Richtung geprägt.

Firth hat den Kotext von Wörtern und Sätzen auf den vier Ebenen der Phonetik und Phonologie, der Morphologie, der Syntax und der Lexik untersucht. Die Untersuchungsbasis bildeten einzelne, situationsgebundene Texte. Heutzutage findet man natürlich eine große Zahl von Sprachhandlungen in Korpora dokumentiert, dies war aber zu Firths Zeiten noch nicht der Fall[38].

Bekannt sind heute noch Firths Arbeiten zur Phonetik und Phonologie und zur Lexik. Die phonetisch-phonologischen Arbeiten sind für die Korpuslinguistik wenig relevant. Interessant sind aber seine Arbeiten zu Wörtern und Kotexten auf der lexikalischen Ebene. Hier spielen die von ihm geprägten Terme *Kollokation* und *Kolligation* eine wichtige Rolle.

Definition 3 (Kollokation). *Innerhalb des Kontextualismus wird unter Kollokation das faktische Miteinandervorkommen zweier oder mehrerer beliebiger Wörter oder lexikalischer Einheiten verstanden. Damit ist keine normative Bewertung hinsichtlich der Korrektheit oder Grammatikalität dieser Wortverbindung verbunden. Der Begriff wird vom späten Firth und einigen seiner Anhänger auf die* Habitualität *des Kovorkommens eingeschränkt. Darunter wird vor allem verstanden, dass die Wortverbindung in den beobachteten Texten wiederholt auftreten muss*[39].

---

[36] Genaueres hierzu in Firth (1991), S. 182.

[37] In vielen linguistischen Arbeiten wird nicht zwischen *Kotext* und *Kontext* unterschieden. Dort wird für beide Bereiche der Ausdruck *Kontext* verwendet, oder es wird zwischen *sprachlichem Kontext* und *nicht-sprachlichem Kontext* unterschieden.

[38] Firth starb im Jahre 1960, das erste größere, digitale Korpus der englischen Sprache wurde 1964 an der Brown University fertiggestellt, vgl. hierzu Kučera und Francis (1967).

[39] Beispiele für Kollokationen finden sich in Fußnote 3.

Die Analyse von Kotext und Kontext linguistischer Einheiten sind für Firth und seine Anhänger der Schlüssel zur Bedeutung dieser linguistischen Einheiten. Bedeutung wird also nicht, wie in vielen anderen Theorien, als eine mentale Disposition von Sprechern oder als eine Struktur, die unabhängig vom Gebrauch existiert, aufgefasst. Damit ist der Kontextualismus eine Gebrauchstheorie der Bedeutung, im Sinne von Wittgensteins berühmter Formel: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"[40]. Firth formuliert dies ganz kontexutalistisch: „You shall know a word by the company it keeps"[41].

Wir werden an späterer Stelle ausführlicher auf Kollokationsanalysen in Korpora eingehen. In diesem, eher theoretischen Zusammenhang ist es wichtig, dass keine Wortverbindung von vornherein ausgeschlossen wird. Jedes Wort kollokiert mit jedem Wort, mit dem es in einer größeren linguistischen Einheit (Satz oder Text) gemeinsam vorkommt. Die Korpusanalyse im Geiste des Kontextualismus ist immer *exhaustiv*, d.h. allumfassend. Der Gebrauch des Korpus durch generative Grammatiker ist, wenn es überhaupt dazu kommt, *selektiv*.

Eine Kombination von lexikalischer Ebene und syntaktischer Ebene im kontextualistischen Rahmen stellt die Kolligation dar.

Definition 4 (Kolligation). *Als Kolligationen werden Paare sprachlicher Einheiten bezeichnet, deren Zusammenhang durch die Bezeichnung ihrer syntaktischen Kategorien und der Beziehungen zwischen diesen Kategorien weiter qualifiziert ist*[42].

Nach dieser Definition ist das Beispiel (13) allein auf Grund des häufigen Kovorkommens als Kollokation aufzufassen, nicht aber als Kolligation. Im Gegensatz dazu ist Beispiel (14) eine Kolligation, da zwischen den beiden Elementen die grammatische Beziehung von Prädikat und Objekt besteht.

(13) und er

(14) Antrag stellen

Mit dem Konzept der Kolligation bekommt die Text- bzw. Korpusanalyse im Rahmen des Kontextualismus ein interpretatorisches Element. Es wird über das reine Erfassen, Auszählen und häufigkeitsbasierte Ordnen von Wortpaaren hinausgegangen. Die gewonnenen Daten werden dadurch *sinnhafter*.

Zusammenfassend lässt sich zur Rolle des Kontextualismus für die moderne Korpuslinguistik sagen:

- Der Kontextualismus ist eine sprachwissenschaftliche Richtung, die linguistische Erkenntnis einzig und allein auf die Analyse des Sprachgebrauchs stützt. Die materielle Basis der linguistischen Untersuchungen, Texte und heutzutage Korpora, werden exhaustiv untersucht. Es werden von vornherein keine Daten ausgeschlossen (etwa, weil sie nicht wohlgeformt wären).

---

[40] Wittgenstein (1967), S. 43.

[41] Firth (1968b), S. 179.

[42] „The study of the collocations in which a word is normally used is to be completed by a statement of the interrelations of the syntactical categories within collocations", Firth (1968a), S. 23.

- Der bedeutendste Beitrag des Kontextualismus für die moderne Korpuslinguistik liegt in der Analyse von Wortverbindungen. Dabei dominiert der syntagmatische Aspekt, das gemeinsame Vorkommen der Wörter in einer größeren linguistischen Einheit, bei weitem den paradigmatischen Aspekt, der im Kontextualismus auch eine Rolle spielt[43]. Wortverbindungen können, je nach dem Status der Interpretation, als *Kollokationen* oder als *Kolligationen* bezeichnet werden.
- Die Analyse von Korpora im Geiste des Kontextualismus hat vor allem im Bereich der Lexikographie und Lexikologie, in der Übersetzungswissenschaft, für den Fremdsprachunterricht und als Basis von sprachkritischen Untersuchungen bedeutende Leistungen ermöglicht.

Generative Grammatik und Kontextualismus unterscheiden sich, wie wir gesehen haben, hinsichtlich der Auffassung ihres Untersuchungsgegenstandes, hinsichtlich dessen, was als sprachliche Daten von Relevanz für die Bildung abstrakter und generalisierter Aussagen über den Gegenstand ist, und dementsprechend auch hinsichtlich der Verwendung von linguistischen Korpora. Ein Austausch zwischen diesen beiden großen Strömungen in der modernen Linguistik fand bisher kaum statt[44]. Erkenntnisse etwa über das kollokative oder funktionale Spektrum lexikalischer Einheiten werden von generativen Grammatikern als trivial und für eine ernsthafte Sprachtheorie irrelevant abgetan. Auf der anderen Seite werden von den Kontextualisten theoretische Aussagen der generativen Grammatiker als unbegründet, da empirisch nicht fundiert oder gar von jeglicher empirischer Basis isoliert und damit empirisch nicht falsifizierbar abgetan.

Wir wollen mit diesem Buch den spezifischen Beitrag von Korpusdaten für alle Arten linguistischer Forschung, für die die beiden dargestellten gegensätzlichen Strömungen stehen, darstellen. Der Beitrag ist natürlich ein jeweils verschiedener, wie die Ausführungen dieses Kapitels zeigen. Die Verwendung von Korpora öffnet aber interessante Wege für die linguistische Forschung insgesamt. Dies wollen wir in den folgenden Kapiteln an einigen, für den heutigen Stand der Forschung typischen Beispielen zeigen.

Im folgenden, abschließenden Abschnitt werden die verschiedenen Ansätze korpusbezogener sprachwissenschaftlicher Forschung überblicksartig dargestellt.

2.4 Korpusbasierte Ansätze

Die in den letzten Abschnitten beschriebenen methodischen Begriffspaare *Empirismus/Rationalismus* und *Deduktion/Induktion* gliedern die folgende Übersicht, siehe auch

---

43 Die Bezeichnungen *syntagmatisch* und *paradigmatisch* gehen auf die Sprachtheorie von Louis Hjelmslev zurück, der sich hier an Ferdinand de Saussure anlehnt, vgl. Hjelmslev (1974). Sprachelemente, die gemeinsam in größeren linguistischen Einheiten vorkommen, stehen in einer syntagmatischen Beziehung zueinander (z.B. ... *Antrag ... stellen*). Sprachelemente, die sich in Kotexten gegenseitig ausschließen und gegeneinander ersetzt werden können, stehen in einer paradigmatischen Beziehung zueinander. Ein Beispiel für eine paradigmatische Beziehung ist die Synonymie, z.B. von *Apfelsine* und *Orange*.

44 Was nicht heißt, dass ein solcher Austausch gar nicht stattfand. Ein Beleg hierfür sind die Arbeiten, die im Tübinger Sonderforschungsbereich 441 *Linguistische Datenstrukturen* entstanden sind, der von 1999 bis 2008 gefördert wurde, ein anderer die bei McEnery et al. dokumentierten Debatten, vgl. McEnery et al. (2006), Abschnitt B2.

Tabelle 1. Wir unterscheiden drei Ansätze in der Korpusanalyse: den korpusbasierten, quantitativen Ansatz, den korpusbasierten, quantitativ-qualitativen Ansatz und den korpusgestützten Ansatz.

	Korpusbasiert quantitativ	**Korpusbasiert, quantitativ und qualitativ**	**Korpusgestützt**
Ansätze	Latent-semantische Analyse N-Gramm Analyse	Kookkurrenz und Kollokation semantische Prosodie	Wortstellungsphänomene
Theoretischer Rahmen	(nicht anwendbar)	Kontextualismus (Firth)	Strukturalismus (Saussure) / Generative Grammatik (Chomsky)
Erkenntnistheoretischer Ansatz	Extrem empiristisch	Gemäßigt empiristisch	Rationalistisch
Personen	Landauer – Jelinek	Sinclair, Teubert, Heringer	Fillmore, Aarts, Oostdijk, Reis, Meurers
Eingabe	Korpus in Rohform	Korpus in Rohform	Linguistisch annotiertes Korpus oder Belegsammlung
Ausgabe	Text-Term Matrizen – N-Gramme mit Frequenzen	Kollokator-Kollokant-Paare mit Kennziffern	Belegsätze
Interpretation	Keine	Ja, von den Belegen ausgehend	Ja, von den theoretischen Aussagen ausgehend
Primäre linguistische Domäne	Statistische Sprachmodelle	Semantik	Syntax
Anwendungsgebiet	Informationserschließung, Verarbeitung gesprochener Sprache	Lexikographie, Fremdsprachunterricht, Übersetzungswissenschaft	Theoretische Linguistik

Tabelle 1: Ansätze in der Korpuslinguistik

2.4.1 Der korpusbasierte, quantitative Ansatz

Bei diesem Verfahren werden auf der Grundlage von rohen, also nicht linguistisch annotierten, Korpora quantitative Daten extrahiert. Diese quantitativen Daten können qualitativ interpretiert werden, dies ist aber für den erfolgreichen Einsatz dieser Verfahren nicht notwendig. Typische Kennziffern einer quantitativen Korpusanalyse sind:

- die absolute Häufigkeit, mit der eine Zeichenkette[45] in einem Text / Korpus vorkommt;
- die relative Häufigkeit[46], mit der eine Zeichenkette in einem Text / Korpus vorkommt;
- der Rangplatz, den eine Zeichenkette auf Grund ihrer Häufigkeit einnimmt (z.B. *sich* ist das zehnthäufigste Wort = das Wort *sich* hat den Rangplatz 10);
- die Distribution eines Wortes, gemessen als die Häufigkeit des Vorkommens dividiert durch die Zahl der Texte des Korpus, in denen das Wort vorkommt;
- Häufigkeiten von Sequenzen – beschrieben als *n-Gramme* – in Texten;
- semantische Ähnlichkeit von Wörtern, gemessen an der Häufigkeit ihres Kovorkommens oder gemeinsamen Vorkommens mit weiteren Wörter (s. Exkurs).

Diese Verfahren werden vor allem im Bereich des Information Retrieval und weiterer texttechnologischer bzw. computerlinguistischer Anwendungen, z.B. der Erkennung und Extraktion von Fachtermen, verwendet. Da sie keine genuin korpuslinguistischen Instrumente sind, gehen wir nur in einem Exkurs auf sie ein.

Exkurs: Quantitative Verfahren im Information Retrieval

Das Ziel des Information Retrieval ist es, auf die Anfrage eines Benutzers die Dokumente zu finden und zu präsentieren, die vermutlich die vom Benutzer gesuchten Informationen enthalten. Sie alle kennen dies von den Suchmaschinen des World Wide Web. Ein Problem, das die Suchergebnisse negativ beeinflusst, ist, dass sehr oft die Wörter der Suchanfrage in Dokumenten nicht vorhanden sind, obwohl formähnliche oder bedeutungsähnliche Wörter vorkommen. Würden diese ebenfalls als Treffer erkannt, dann würden auch diese, für die Anfrage relevanten, Dokumente gefunden. Wir wollen hier kurz den *n-Gramm*-Ansatz für das Auffinden formähnlicher Wörter und auf die latente semantische Analyse für das Auffinden bedeutungsähnlicher Wörter eingehen.

N-Gramme: Vorkommenshäufigkeiten von *n-Grammen* linguistischer Einheiten können dazu verwendet werden, formähnliche Wörter in Anfrage und Text aufeinander abzubilden. Es kann sich dabei um Folgen von 1, 2, 3, ... *n* Phonemen, Graphemen etc. handeln. Nehmen wir an, dass in einem Text *Operationen am offenen Herzen* vorkommt. In der Suchanfrage wird der Term *Herzoperation* verwendet. Bei einfachem Abgleich der Wörter würde das Dokument, das doch immerhin relevant erscheint, nicht gefunden. Beide Zeichenketten haben aber acht Trigramme gemeinsam *_He, Her, erz, per, era, ati, tio, ion*, also ca. 90 Prozent der Trigramme der kürzeren Zeichenkette. Das n-Gramm-Verfahren ist eine Möglichkeit, der Schreibvarianten bei vielen Wörtern und Termen Herr zu werden. N-Gramm-Modelle werden ausführlich in Jurafsky und Martin (2008), Kap. 4 behandelt.

Latent-semantische Analyse: Um semantisch ähnliche Wörter in Anfrage und Dokumenten aufeinander abzubilden, wird aus dem Vorkommen von Termen in Dokumenten deren Ähnlichkeit bestimmt. Ist ein Term in der Anfrage einem Term in einem Dokument semantisch ähnlich, dann steigert dies die Relevanz dieses Dokumentes für die Anfrage.

---

[45] Zeichenketten sind das Ergebnis der Segmentierung von Texten, s. Kapitel 4. Da Texte meist in Wörter zerlegt werden, könnte man stattdessen auch von *Wörtern* sprechen. *Zeichenkette* ist aber der präzisere Ausdruck.

[46] Siehe hierzu Kapitel 6.

Eine Matrix ist eine Tabelle mit Zeilen und Spalten. Im Fall einer Term-Dokument-Matrix nimmt jedes Wort eine Zeile ein und jeder Text eine Spalte. Die folgenden, sehr kurzen Texte:

(15) Miliz verhaftet Terroristen nach Anschlag (= Text 1)

(16) Terroristen verüben Anschlag (= Text 2)

können wie links als Matrix repräsentiert werden (vgl. rechts in explizitem Tabellenformat):

Miliz	(1,0)
verhaftet	(1,0)
Terroristen	(1,1)
nach	(1,0)
Anschlag	(1,1)
verüben	(0,1)

Wort	Text 1	Text 2
Miliz	1	0
verhaftet	1	0
Terroristen	1	1
nach	1	0
Anschlag	1	1
verüben	0	1

Hat man viele Dokumente vorliegen, und damit viele verschiedene Wortformen, dann entsteht eine sehr große Matrix (mit m Zeilen für m Wortformen und n Spalten für n Dokumente). Die Matrix enthält viele Leerstellen, d.h. Nullvorkommen, da die meisten Wörter in den meisten Texten nicht vorkommen.

Die sogenannte Singulärwertzerlegung als mathematische Operation über Matrizen[47] bietet die Möglichkeit, solche großen Matrizen auf einige Hundert Dimensionen (= Zeilen und Spalten) zu verkleinern, bei optimaler Erhaltung der in ihnen kodierten Informationen.

Intuitiv lässt sich der Effekt dieser Verkleinerung wie folgt beschreiben: Ein Term, der in einem bestimmten Text nicht vorkommt, dafür aber gemeinsam (in anderen Texten) mit vielen Termen vorkommt, die für diesen Text relevant sind, erhält Gewicht auch für diesen Text, in dem er, wie gesagt, gar nicht vorkommt. Terme wiederum, die in diesem Text zwar vorkommen, aber keine enge Beziehung zu den anderen, für diesen Text relevanten Termen haben, werden heruntergewichtet. So kann es sein, dass der Term *Nabe* für einen Text über Fahrräder ein relativ hohes Gewicht erhält, obwohl er gar nicht darin vorkommt, wohl aber oft in der Nachbarschaft von *Speiche, Felge* etc.

Der Effekt dieser Terme und Dokumente verknüpfenden Matrix ist also, dass auch der Text über Fahrräder als relevant angezeigt wird, obwohl das Wort *Nabe* nicht in ihm vorkommt.

Deshalb eignet sich dieses Verfahren für die Informationserschließung, wo es um die Ähnlichkeit von Suchanfrage und Zieldokument geht. Dort ist dieses Verfahren unter dem Namen *Latent-semantische Indexierung* bekannt.

Es eignet sich aber auch für die Ermittlung der semantischen Ähnlichkeit von Wörtern. Dort trägt das Verfahren den Namen *Latent-semantische Analyse*. In unserem Zusammenhang ist es wichtig, dass bei diesen Verfahren weder eine linguistische Analyse der Textkorpora noch eine linguistische Analyse der resultierenden Daten erfolgt[48].

---

[47] Zum mathematischen Hintergrund vgl. Berry et al. (1999).

[48] Zur Einführung in die latent-semantische Analyse empfiehlt sich die Lektüre von Landauer und Dumais (1997) (theoretischer Hintergrund) und Landauer et al. (1998) (Anwen-

2.4.2 Der korpusbasierte, quantitativ-qualitative Ansatz

Dieser Ansatz ist dem soeben beschriebenen sehr ähnlich. Wie wir weiter oben gezeigt haben, wird auch in diesem, dem Kontextualismus verpflichteten Forschungsprogramm das Korpus exhaustiv analysiert. Es bildet die ausschließliche Basis für linguistische Untersuchungen, andere Quellen wie Experimente und Sprecherbefragungen werden ausgeschlossen. Die Beobachtung des Sprachgebrauchs bildet die Hauptquelle der linguistischen Erkenntnis.

Ein wichtiger Unterschied zu dem vorherigen Ansatz ist es, dass hier die Daten, die aus Korpora abgeleitet sind, nicht uninterpretiert bleiben. Zur Interpretation der Daten werden, zumindest bei einigen Vertretern des Kontextualismus, grammatische Kategorien herangezogen, die nicht aus den Daten selber abgeleitet wurden. Auch hat der Kontextualismus den Anspruch, etwas über das Sprachsystem (einer Einzelsprache) auszusagen. Wie dies durch Generalisierung der Beobachtungsdaten gelingen kann, das wird allerdings nicht thematisiert. Wir werden in einem späteren Kapitel auf korpusbasierte Untersuchungen im Rahmen dieses Forschungseinsatzes eingehen.

Der größte Nutzen dieser Art von Korpuslinguistik konnte bisher in der Lexikographie, in der Übersetzungswissenschaft und für den Fremdsprachunterricht erzielt werden[49]. Auch für sprachkritische Untersuchungen erwies sich der Ansatz als fruchtbar.

2.4.3 Der korpusgestützte Ansatz

Sprachtheorien im Geiste der generativen Grammatik berücksichtigen Korpusdaten, wenn überhaupt, dann nur als zusätzliche Quelle der Evidenz. Wenn Korpora herangezogen werden, dann sind sie nicht als Ganzes interessant. Es wird in ihnen gezielt nach relevanten (meist syntaktischen) Konstruktionen gesucht, um Voraussagen, die aus einer Theorie folgen, zu bestätigen oder zu widerlegen. Dabei ist der Status oder Wert solcher *e-sprachlichen* Belege umstritten.

Erschwerend kommt hinzu, dass in den Korpora oft nach relativ komplexen Konstruktionen aus lexikalischen und grammatischen Elementen, die hohe Variabilität haben können, gesucht werden muss. Dem sind die meisten Korpusabfragesprachen nicht gewachsen. Die Benutzung eines Korpus gleicht also oftmals der Suche nach einer Nadel im Heuhaufen. Auch dies trägt sicher nicht zur Akzeptanz von Korpora in der generativen Grammatik bei[50].

2.5 Weiterführende Literatur

Der theoretische Hintergrund der modernen Korpuslinguistik wird leider nur sehr selten thematisiert. Ein paar Seiten hierzu finden sich bei McEnery und Wilson (1996), Kapitel 1. Paprotté geht in zwei Aufsätzen etwas genauer auf diese Frage ein (Paprotté,

---

dungen). Eine hervorragende, wenn auch recht anspruchsvolle Einführung in das Thema ist Widdows (2004). Mittlerweile gibt es auch eine recht erfolgreiche, an Wikipedia-Daten entwickelte „Explicit Lexical-Semantic Analysis", vgl. Grabilovich und Markovitch (2007).

[49] Vgl. z.B. die Fallstudien in Tognini-Bonelli (2001).

[50] In den letzten Jahren ist hier jedoch ein Wandel zu beobachten, s. z.B. Bresnan et al. (2007) und Beiträge der Tagung *Linguistic Evidence*, vgl. z.B. Kepser und Reis (2008).

1992, 1994). Ein lebendiges Bild der linguistischen Szene und ihrer Kämpfe vermittelt das Buch „The linguistic wars" (Harris, 1995). Dieses Buch eignet sich aber eher als Bettlektüre denn als Referenzwerk. Der frühe Chomsky wird in einem Handbuchartikel von Karlsson (2008) analysiert. Über die Positionen des Kontextualismus gibt Tognini-Bonelli (2001) Auskunft. Von diesem Buch haben wir die Unterscheidung in korpusbasierte und korpusgestützte Untersuchungen übernommen. Die Autorin argumentiert allerdings ganz aus der Sicht des Kontextualismus und ist insofern anderen Ansätzen gegenüber nicht immer ganz fair. In einem von Svartvik herausgegebenen Band eines hochkarätigen Symposiums (Svartvik, 1992) werden methodische Fragen reflektiert. Einige dort versammelte Aufsätze aus diesem Band sind aus dieser Perspektive besonders ergiebig (vor allem Fillmore, 1992; Chafe, 1992; Halliday, 1992; Leech, 1992).

McEnery und Hardie (2012) stellen den korpusgestützten ('corpus-based') und den korpusbasierten ('corpus-driven') Ansatz ausführlich gegenüber und widersprechen den Ansichten von Tognini-Bonelli grundsätzlich. Schon deshalb ist dieses Buch im Anschluss an letzteres sehr lesenswert.

Schließlich sei noch auf die von Sampson und McCarthy (2004) herausgegebene Aufsatzsammlung hingewiesen, in der viele der hier angeschnittenen Fragen ausführlicher behandelt werden.

2.6 Aufgaben

1. Welche der folgenden Aussagen sind empirisch begründet und welche rationalistisch:
   a) Der Satz *Wo sollen wir treffen?* ist ungrammatisch.
   b) Ein Satz wie *Wo sollen wir treffen?* resultiert aus einem typischen Fehler englischer Lerner des Deutschen bei der Verwendung des Verbs *treffen*.
   c) Der Kopf einer Nominalphrase ist das Nomen.
   d) Instruktive Texte (z.B. Kochrezepte) enthalten überdurchschnittlich viele Befehlsformen.
   e) Es gibt im Deutschen fünfzehn Dialekte.
   f) In der Tiefenstruktur des deutschen Satzes steht das finite Verb am Satzende. Bei einigen Satztypen wird es bei der Realisierung der Oberflächenstruktur an Zweitposition verschoben.
   g) Kollokationen sind Paare von Wörtern, die überdurchschnittlich häufig miteinander vorkommen.
   h) *Hartes Leben* ist eine Kollokation.

   Für die Beantwortung welcher Frage benötigen Sie ein Korpus?
2. Welche Gründe sprechen dagegen, von Performanzdaten auf die Kompetenz der Sprecher zu schließen? Sind die Probleme, die solche Schlüsse mit sich bringen, dadurch behebbar, dass man ein größeres oder variantenreicheres Korpus wählt?
3. Stellen Sie die Unterschiede zwischen dem korpusbasierten Forschungsansatz und dem korpusgestützten Forschungsansatz dar. Für welche Arten linguistischer Untersuchungen eignet sich der korpusbasierte Ansatz eher, für welche der korpusgestützte?

3 Der Stein der Weisen? — Linguistische Korpora

Am Ende dieses Kapitels kennen Sie die wichtigsten Merkmale linguistischer Korpora und wissen, was diese von anderen linguistischen Datensammlungen unterscheidet. Sie können die drei Datenebenen von Korpora benennen und wissen, welche Probleme man auf den verschiedenen Ebenen berücksichtigen muss. Sie können schließlich im Rahmen Ihrer eigenen Untersuchung Antworten auf drei schwierige methodologische Fragen formulieren: Wie verhält sich mein Korpus zum Gegenstand, den ich eigentlich untersuchen will? Was mache ich, wenn ich im Korpus etwas nicht finde, was ich suche und beschreiben möchte, und umgekehrt: Was mache ich, wenn ich etwas finde, was es nach einer bestimmten Sprachtheorie eigentlich gar nicht geben dürfte? Sie sind nun für eigene linguistischen Untersuchungen an Korpora gut gerüstet!

3.1 Definition und Abgrenzung

In diesem Abschnitt wollen wir die Definition von *Korpus* aus der Einleitung weiter präzisieren.

Definition 1 (Korpus). *Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.*

Wenn wir von *linguistischen Korpora* sprechen, dann in dem Sinne, dass es sich

- um Textsammlungen mit kompletten Texten oder zumindest mit sehr großen Textausschnitten

handelt. Außerdem sind linguistische Korpora oft, aber nicht immer

- repräsentativ für den Gegenstand, auf den sie sich beziehen,
- durch Metadaten erschlossen,
- linguistisch annotiert.

Das erste Kriterium qualifiziert Korpora als solche und unterscheidet sie von anderen Sammlungen linguistischer Daten. Die Grenze zwischen Korpora im engeren Sinn und

anderen Datensammlungen ist aber nicht absolut. So werden, wie wir in Abschnitt 7.4 zeigen, große Webkorpora aus urheberrechtlichen Gründen in Sätze zerlegt und diese Sätze nach dem Zufallsprinzip angeordnet. Die einzelnen Texte lassen sich somit nicht mehr rekonstruieren. Dennoch sind diese Korpora im weiteren Sinn dazu verwendbar, um linguistische Theorien oder Hypothesen zu überprüfen, sofern diese sich auf Phänomene beziehen, die man innerhalb der Grenzen eines Satzes beobachten kann.

Die anderen Merkmale zeichnen vor allem größere Korpora aus, nicht aber die vielen kleineren Korpora, die im Rahmen einer einzelnen Untersuchung gebildet wurden.

Die meisten modernen Korpora bestehen aus einer Sammlung vollständiger Texte oder Gespräche, z.B. aus Zeitungsartikeln oder Chatprotokollen. Texte können sehr kurz sein, zum Beispiel bei SMS oder Tweets, deren Länge aus technischen Gründen auf unter 200 Zeichen begrenzt sind[1]. Andere Texte sind sehr lang, zum Beispiel bei Romanen. Wichtig ist es für viele linguistische Untersuchungen, dass der Textausschnitt um ein bestimmtes Phänomen herum groß genug ist, damit zum Beispiel pronominale und kontextuelle Bezüge aufgelöst werden können. Deshalb wird auch meist nicht ein einzelner Satz, sondern eine größere Textsequenz untersucht.

In den sechziger Jahren, als das *Brown Corpus* („A Standard Corpus of Present-Day American English") entstand, war die Digitalisierung und Speicherung vieler langer Texte nicht möglich. Die Ersteller dieses Korpus, Henry Kučera und Nelson Francis, entschieden sich deshalb dafür, von fünfhundert Texten unterschiedlicher Textsorten jeweils einen Ausschnitt von 2000 Wörtern aufzunehmen. Auch wenn viele der aufgenommenen Texte deshalb nicht vollständig sind, repräsentieren sie einen hinreichend großen Ausschnitt an fortlaufendem Text, und die Leistung ist für die damalige Zeit beachtlich[2].

Auf die weiteren Kriterien, Metadaten und Repräsentativität, gehen wir in den folgenden Abschnitten dieses Kapitels ein. Der linguistischen Annotation sind zwei eigene Kapitel gewidmet (Kapitel 4 und 5).

Die genannten Kriterien bzw. Anforderungen an linguistische Korpora sind geeignet, diese von anderen Sammlungen sprachlicher Daten abzugrenzen.

3.1.1 Korpora für nicht-linguistische Zwecke

Einige Korpora, wie das *Corpus Iuris Civilis* und das *Corpus Iuris Canonici* versammeln juristische Texte, deren Erforschung vor allem für Rechtshistoriker von Interesse ist. Hinter Namen wie *Corpus Christianorum* verbergen sich Sammlungen von Texten der christlichen Kirchengeschichte.

Auf dem weltlichen Gebiet gibt es große Archive literarischer Texte, deren Urheberrecht verfallen ist[3]. Das bekannteste Projekt ist das *Projekt Gutenberg*, in dem Freiwillige klassische literarische Texte elektronisch erfassen[4]. Auch das Projekt *digitale Bibliothek*[5] fällt unter diese Kategorie, die wir *Textarchiv* nennen wollen. Texte aus der Zeit von

---

1 Eine von Louvain ausgehende Initiative baut in mehreren Ländern, u.a. in der Schweiz, zurzeit SMS-Korpora auf, s. `http://www.sms4science.ch`.

2 Vgl. Kučera und Francis (1967), Introduction.

3 Die Urheberrechte eines Autors bzw. seiner Erben verfallen siebzig Jahre nach dem Tod des Autors, jedenfalls nach deutschem Recht.

4 Die Adresse des deutschen Gutenberg-Projekts lautet: `http://gutenberg.spiegel.de/`.

5 Die Adresse lautet: `http://www.digbib.org/`.

1650 bis 1900 werden zurzeit in größerem Umfang im Projekt *Deutsches Textarchiv* an der Berlin-Brandenburgischen Akademie der Wissenschaften digitalisiert und der Forschung zur Verfügung gestellt[6]. Solche Texte sind selbstverständlich auch für linguistische Untersuchungen brauchbar und nützlich, zum Beispiel wenn man die Existenz oder Verbreitung eines bestimmten sprachlichen Phänomens in einem bestimmten Sprachstadium nachweisen möchte. Der ursprüngliche Zweck dieser Archive ist aber vor allem, die dort digital erfassten und gespeicherten Texte zu sichern, auf einem Medium, dass hoffentlich beständiger ist als Papier.

3.1.2 Linguistische Belegsammlungen

Darüber hinaus gibt es zu lexikographischen und linguistischen Zwecken angelegte Belegsammlungen.

- Die bekannteste Belegsammlung ist sicher die Duden-Sprachkartei[7]. In ihr wurden früher mechanisch, heute elektronisch Belege zu den Wörtern erfasst, die in den Wörterbüchern der Duden-Reihe gebucht sind.
- In der *Wortwarte*[8] sind gut 60 000 zwischen 2000 und heute neu geprägte Wörter, jeweils mit mindestens einem Beleg, erfasst. Auch das Institut für deutsche Sprache verfügt über eine elektronische Kartei mit Neologismen[9].

Dies sind lexikalisch orientierte Belegsammlungen. Ihnen vergleichbar sind Sammlungen von Belegen syntaktischer Muster, wie die Testsatzsammlung von Stefan Müller[10], die von Istvan Batori und Martin Volk aufgebaute *Grammatiktestumgebung*[11], die Sammlung *suboptimaler syntaktischer Strukturen* von Wolfgang Sternefeld u.a.[12] sowie die *Collection of Distributionally Idiosyncratic Items* (CoDII)[13]. Der Vorzug dieser Satzsammlungen ist, dass sprachliche Phänomene dokumentiert werden können, die in Korpora nur selten oder gar nicht vorkommen[14].

Bei vielen linguistischen Untersuchungen wird mit Belegsammlungen gearbeitet, die für den Zweck dieser Untersuchung aus den verwendeten Korpora extrahiert wurden. Diese Belegsammlungen bilden oft die Basis qualitativer Analysen, während das gesamte Korpus vor allem für quantitative Analysen herangezogen wird. Wir werden in Kapitel 8 solche Untersuchungen im Detail beschreiben.

---

6 `http://www.deutschestextarchiv.de`.

7 Vgl. Scholze-Stubenrecht (2002). Diese Kartei nennen wir hier als ein Beispiel unter vielen. Auch anderen große Wörterbuchprojekte wie das *Deutsche Wörterbuch*, gegründet von den Gebrüdern Grimm, oder das *Oxford English Dictionary* verfügen über große Belegsammlungen.

8 Adresse: `www.wortwarte.de`.

9 Vgl. Herberg et al. (2004), S. XVI f.

10 Adresse: `http://hpsg.fu-berlin.de/Software/TS/`.

11 Näheres hierzu unter `http://www.zora.uzh.ch/19140/` sowie in Volk (1995).

12 Siehe `www.tcl-sfs.uni-tuebingen.de/~kepser/papers/tlt2004.pdf/`.

13 Siehe `http://www.lingexp.uni-tuebingen.de/sfb441/a5/codii/info-bu-de.xhtml`.

14 Generell zu diesem Thema s. Bergh und Zanchetta (2008).

3.1.3 Ist das World Wide Web ein Korpus?

Eine Frage, die immer wieder gestellt wird, lautet: Ist das World Wide Web ein Korpus? Einige Korpuslinguisten, die diese Frage aufgeworfen haben, beantworten diese Frage mit „Ja“[15]. In der Tat kann man in World Wide Web große Mengen authentischer Texte in allen möglichen Sprachen finden[16]. Wenn es lediglich um die Datenmenge geht, dann ist das World Wide Web eine gute Quelle für linguistische Untersuchungen. Es muss aber zum Beispiel das Problem gelöst werden, deutschsprachige Texte zu finden[17], d.h. diese von den Texten oder Textteilen in anderen Sprachen zu trennen. Dann ist es keineswegs leicht, fortlaufenden Text von textähnlichen Artefakten wie Tabellen oder Teilen von Programmcode zu trennen. Schließlich gibt es kaum Daten über die Herkunft, den Entstehungszeitpunkt oder die Autorschaft von Texten. Solche so genannten *Metadaten*, auf die wir im nächsten Abschnitt genauer eingehen werden, sind im World Wide Web in seinem heutigen Zustand kaum zu finden[18].

Für viele sprachstatistische Untersuchungen liegt der Vorrang in der Verfügbarkeit großer Datenmengen, egal welcher Herkunft. Wer genauer beschriebene Daten für qualitative linguistische Untersuchungen benötigt, kann das World Wide Web als Quelle nutzen, sollte aber einiges an Aufwand für die Bereinigung und Beschreibung dieser Daten einplanen.

Problematisch wird die Benutzung des WWW als Textkorpus, wenn man nach seltenen Konstruktionen sucht oder nach Beispielen, über deren Grammatikalität man im Unklaren ist. Man findet dann tatsächlich oft solche Beispiele, aber wenn man genauer hinsieht, merkt man, dass sie in online verfügbaren linguistischen Texten auftreten und im Text dann oft als ungrammatische Beispiele angeführt und diskutiert werden. Eine andere Quelle der Unsicherheit sind Texte, die von Nicht-Muttersprachlern der jeweiligen Sprache verfasst wurden.

Die oben gestellte Frage sollte wie folgt verändert werden: Lassen sich Teile der im WWW verfügbaren Daten sinnvoll als Korpora für linguistische Untersuchungen nutzen? Als solche kann man sie momentan mit „Ja, aber...“ beantworten. Dass dieses Thema Linguisten beschäftigt, darauf deutet u.a. die Publikation von zwei Sammelbänden zu diesem Thema hin (Baroni und Bernardini, 2006; Hundt et al., 2007). Darin geht es sowohl um organisatorische und methodische Probleme bei der Nutzung von Web-Daten für linguistische Korpora als auch um konkrete Projekte. Serge Sharoff schließlich zeigt

---

15 „The answer to the question ‚Is the web a corpus‘ is yes.“ (Kilgarriff und Grefenstette, 2003, S. 334).

16 Nach einer gut begründeten Schätzung von Grefenstette und Kilgarriff aus dem Jahr 2000 fand man damals im World Wide Web Texte im Umfang von gut 3 Milliarden Wörter. Selbst für Sprachen wie Baskisch konnte man von einem Volumen von weit über 50 Millionen Textwörter ausgehen, vgl. Kilgarriff und Grefenstette (2003), S. 337ff. und Tabellen 2 und 3.

17 Genauer, und noch schwieriger: In einem Korpus der deutschen Sprache sollten nur Texte von Muttersprachlern vertreten sein. Dies ist bei der Zusammenstellung eines Webkorpus beim heutigen Stand der Texte sehr schwer, wenn nicht gar unmöglich zu kontrollieren.

18 Diese Situation beginnt sich aber zu bessern, was mit den Bemühungen innerhalb der WWW-Gemeinschaft zu tun hat, mittels verschiedener Datenbeschreibungssprachen wie XML, RDF, RSS etc., die die Beschreibung von Texten, die ins Web gestellt werden sollen, erleichtern.

in einem Aufsatz (Sharoff, 2006) und auf einer Webseite[19], wie man mit relativ wenig Aufwand ein Korpus aus Webtexten einer bestimmten Sprache zusammenstellen kann.

Im folgenden Abschnitt werden wir auf die verschiedenen Typen von Daten eingehen, aus denen ein linguistisches Korpus bestehen kann.

3.2 Primärdaten und Metadaten

Im einfachsten Fall besteht ein Korpus lediglich aus den Daten, die in diesem Korpus erfasst wurden, den Primärdaten. In einem guten Korpus findet man außerdem Daten, die über die Herkunft dieser Äußerungen bzw. Texte und über einiges mehr Auskunft geben. Wir werden diese Daten *Metadaten* nennen. Schließlich wurden und werden Korpora linguistisch annotiert, die linguistischen Einheiten werden also mit ihren linguistischen Beschreibungen verbunden. Dies werden wir *Annotation* nennen[20].

3.2.1 Primärdaten

Aus der Tatsache, dass Texte in ein Korpus aufgenommen werden, könnte man zunächst schließen, dass man hier ein getreues Abbild dieser Texte hat. Dies ist nicht der Fall und man sollte sich dessen bei der Benutzung eines Korpus immer bewusst sein. Wir nennen hier nur einige Beispiele, die für linguistische Untersuchungen problematisch werden könnten:

- Es ist wohl am offensichtlichsten, dass die Transkriptionen gesprochener Äußerungen immer Vereinfachungen und damit Interpretationen sind[21]. Die für das Sprechen so wichtigen begleitenden parasprachlichen Signale, wie z.B. Gestik oder Mimik, und auch einige sprachliche Signale wie die Tonhöhe sind nur schwer oder auch gar nicht in das geschriebene Medium zu übertragen. Im Zweifelsfall helfen nur Videoaufnahmen, die Bild und Ton einer Äußerungssituation exakt wiedergeben.
- Eigenschaften von Texten, die nicht sprachlich motiviert sind, wie die Worttrennung am Zeilenende[22] oder Schrifttyp, Schriftschnitt und Schriftgröße des Originaltextes werden bei der Übernahme eines Textes in ein Korpus oft stillschweigend ausgeblendet. Dies kann vereinzelt zu Problemen führen, z.B. wenn nicht rekonstruiert werden kann, ob ein Strich am Zeilenende ein Trennstrich oder zusätzlich ein Bindestrich sein soll (Ist *4-(Umbruch)türig* auf *4türig* oder auf *4-türig* zurückzuführen?). Solche Mehrdeutigkeiten als Ergebnis des Ausblendens der Silbentrennung am Zeilenende sind eher selten, aber nicht auszuschließen.
- Will man die Informationsverteilung in Texten untersuchen, dann kann es wichtig sein zu wissen, dass ein Teil eines Textes in der Originalquelle auf der Titelseite, der Rest des Textes im Zeitungsinneren gedruckt wurde. Diese Aufteilung kann dazu führen, dass der Text so aufgebaut wurde, dass der Leser zum Weiterlesen des Textes im Heftinneren angeregt wird. In einem guten Korpus ist diese kontextuelle Information vermerkt, man kann aber nicht immer damit rechnen.

---

[19] Die Adresse: `http://corpus.leeds.ac.uk/internet.html`.

[20] Vgl. ausführlich Kapitel 4.

[21] Vgl. hierzu Schmidt (2005) sowie Draxler (2008), S. 173f.

[22] Die Worttrennung am Zeilenende sollte allerdings sprachlichen Normen folgen.

In vielen Fällen wird man nicht umhin können, sich das Original des Textes anzusehen und so die nicht-sprachlichen Informationen zu erschließen. In einigen industriellen Archivierungsprojekten ist aus juristischen Gründen üblich, nur das gescannte *Bild* eines Dokumentes aufzuheben, nicht jedoch den durch OCR-Software[23] digitaliserten Text[24].

3.2.2 Metadaten

Metadaten sind *Daten über Daten*. Genauer, und in unserem Zusammenhang passender: Metadaten sind Daten, die verschiedene Aspekte einer Informationsressource beschreiben. Die Informationsressource kann z.B. ein Text sein, eine Textsammlung, eine Tonaufnahme oder ein Video. Die Aspekte, unter denen eine Informationsressource beschrieben werden kann, sind z.B. ihr Inhalt, das Trägermedium, die Art der Kodierung, die Autoren und andere bei der Produktion beteiligte Personen, der Zeitpunkt der Entstehung[25].

Metadaten sind entweder Bestandteil der Daten, die sie beschreiben – dies ist zum Beispiel bei den Titelseiten eines Buches der Fall – oder sie werden von den beschriebenen Daten getrennt erfasst und gespeichert – wie zum Beispiel bei Karteikarten in Bibliotheken. Man spricht im letzteren Fall von dem Metadatenobjekt als *Stellvertreter* des eigentlichen Informationsobjekts. Die Bedeutung von Metadaten ist umso größer, je schwerer zugänglich die Primärdaten sind.

Funktionen von Metadaten

Metadaten erfüllen die folgenden Funktionen:

- Sie dokumentieren vor allem kontextuelle Aspekte der Entstehung und Entwicklung des beschriebenen Objekts. Diese Informationen sind den meisten späteren Benutzern anders nicht zugänglich. Zu den dokumentierten Aspekten etwa eines Textes gehören die Entstehungszeit, die Druck- bzw. Publikationszeit, Publikationsort, beteiligte Personen usw.
- Sie liefern den Schlüssel zu den Primärdaten. Wenn die Filme in einem Filmarchiv mit den entsprechenden Metadaten versehen sind, dann können Sie dort alle Filme recherchieren, in denen Humphrey Bogart und Lauren Bacall zusammen auftraten. Sie mögen Woody Allen als Regisseur, nicht aber seine schauspielerischen Leistungen? Im Archiv sollten sich die Filme finden lassen, an denen Woody Allen mitwirkte, aber nur in der Rolle als Regisseur.

Für Korpuslinguisten spielen natürlich andere Kriterien der zu untersuchenden Daten eine Rolle. Der dokumentierte Entstehungszeitpunkt von Texten (oder Tonaufnahmen) erlaubt es, Teilkorpora zusammenzustellen, die die Sprache einer bestimmten Epoche

---

[23] *OCR* steht für automatische Buchstabenerkennung (‚optical character recognition').

[24] Da es bei historischen Texten noch mehr als bei aktuellen Texten darauf ankommt, Eigenschaften des Schriftbildes, des Seitenlayouts usw. zu erfassen, stellt das *Deutsche Textarchiv* (mehr dazu in Kap. 7) neben den (digitalisierten) Texten auch die Scans der Buchseiten als Digitalisate zur Verfügung. Die Fundstelle zu einem Suchwort wird auf der entsprechenden Seite hervorgehoben, um die Suche auf dieser Seite zu erleichtern.

[25] Unsere Darstellung über Metadaten orientiert sich an Schmidt (2004).

bzw. Sprachstufe dokumentieren (*die deutsche Sprache der Goethezeit, die deutsche Sprache der Wendezeit*, etc.) oder die Sprache einer bestimmten Region (*das Oberschwäbische, die Sprache in der DDR*). Der Fokus kann auf bestimmte Textsorten oder Genres gelegt werden (*die Sprache von Gebrauchsanweisungen, Formen der Höflichkeit in Erpresserbriefen*). Diese und einige andere Merkmale von Texten müssen entsprechend als Metadaten kodiert worden sein, damit solch präzise Definitionen von Teilkorpora möglich sind.

Die Metadaten für digitale Korpora und einzelne Texte, die Bestandteile von solchen Korpora sind, müssen den Umstand berücksichtigen, dass möglicherweise zwei Informationsobjekte beschrieben werden müssen:

1. Das Informationsobjekt, auf das sich die Metadaten direkt beziehen, ist z.B. ein Text in seiner digitalen Form, der sich an einer bestimmten Stelle als Datei auf einem digitalen Datenträger (Festplatte, CD-ROM etc.) befindet, einen bestimmten Namen hat und dessen einzelne Zeichen einer bestimmten Konvention folgend kodiert, also in Bits und Bytes abgebildet wurden.
2. Das Informationsobjekt, aus dem die digitale Datei gewonnen wurde, z.B. durch Abtippen, Einscannen oder Einlesen eines Druckereidatenträgers. Dies kann ein Zeitungsartikel sein, der in einer bestimmten Ausgabe einer Zeitung erschien, ein Text aus einem Kinderbuch, eine Tagebuchseite usw.

Beide Informationsobjekte führen ein getrenntes Dasein, und streng genommen beziehen sich die Metadaten, die wir hier meinen, nur auf das erste Informationsobjekt. Der digitalisierte Text kann zum Beispiel die Abschrift einer Geschichte aus einem Kinderbuch sein, das seitdem in einer neuen Auflage in neuer Rechtschreibung herausgegeben wurde. Es ist deshalb wichtig, in den Metadaten zu einem digitalisierten Text möglichst genau auf die Quelle dieses Textes, das Original, hinzuweisen. Es ist außerdem sinnvoll, in den Metadaten auf die Person hinzuweisen, die den Datensatz erstellt hat. Das sind dann *Meta-Metadaten*.

Wenn Sie ein Korpus benutzen wollen, dann ist es wichtig zu erfahren, ob es Metadaten zu dem gesuchten Korpus gibt, welche Informationen diese enthalten, und ob Sie darauf Zugriff haben. Einige dieser Fragen werden wir in dem Kapitel, in dem wir einige deutsche Korpora beschreiben, individuell beantworten. Noch gibt es keine zentralen Stellen, wie etwa die Bibliotheken und ihre Kataloge, wo Sie separat in Metadaten zu Sprachressourcen suchen können. Einige Institutionen dieser Art sind aber im Aufbau[26].

3.2.3 Metadaten für Ihr eigenes Korpus

Wenn Sie selber ein Korpus aufbauen wollen, dann stellen sich die folgenden Fragen: Sollten Sie Ihre Daten mit Metadaten anreichern? Wenn ja, welches Format ist dafür am besten geeignet?

Die erste Frage lässt sich nicht generell beantworten. Sie sollten Ihre Korpusdaten mit Metadaten anreichern, wenn das Folgende für Sie bzw. Ihre Daten zutrifft:

---

[26] Zu nennen sind hier die *Open Language Archives Community* (OLAC); Adresse: `http://www.language-archives.org/`, das zentrale Repositorium für Sprachressourcen und sprachtechnologische Werkzeuge des CLARIN-Projekts (`http://www.clarin.eu/vlo`) und das LAUDATIO-Repositorium (`http://www.laudatio-repository.org/repository/`).

- Sie sind nicht die einzige Person, die Daten zum geplanten Korpus beiträgt;
- Sie möchten Forschungsergebnisse mit dem Korpus in einer Weise belegen und dokumentieren, die für die Leser Ihrer Arbeiten nachvollziehbar und nachprüfbar ist;
- Sie werden voraussichtlich nicht der einzige Benutzer der Daten sein;
- Sie möchten die Möglichkeit haben bzw. schaffen, Teile der Daten nach bestimmten Kriterien auszuwählen.

Je mehr dieser Punkte auf Ihre Pläne zutreffen, um so dringender ist dazu zu raten, dass Sie sich mit der Erstellung von Metadaten vertraut machen. Wenn Sie eine kleine Menge von Daten für eine begrenzte Untersuchung sammeln, die Überprüfbarkeit Ihrer Thesen an einzelnen Texten nicht relevant ist und Sie diese Texte auch nicht anderen Forschern zur Verfügung stellen können oder wollen, dann ist der Aufwand, den Sie in die Erstellung von Metadaten stecken würden, wahrscheinlich zu hoch. Aber bedenken Sie: Es ist schwierig bis unmöglich, Daten nachträglich mit Metadaten zu versehen.

Standards für Metadaten

Die Interoperabilität zwischen Korpora und den für ihre Analyse benötigten sprachtechnologischen Werkzeugen verbessert sich, wenn beide mit ausreichenden und geeigneten Metadaten beschrieben sind. Mit *Interoperabilität* ist Folgendes gemeint a) der Austausch von Korpora und die Bildung von Teilkorpora nach textexternen Kriterien, auch über die Grenzen einzelner Korpora hinweg und b) die Nutzung weit verbreiteter sprachtechnologischer Werkzeuge für beliebige Korpusdaten. Bei der Vergabe von Metadaten sollte man sich dabei an Standards halten. Für (linguistische) Korpora wurden verschiedene Metadaten-Standards entwickelt, die wir hier kurz vorstellen wollen.

- *Dublin Core Metadata Initiative*[27]. Dublin Core (DC) ist ein Schema zur Beschreibung von elektronischen Ressourcen. Mittlerweile hat sich DC zu einem internationalen Übereinkommen über eine Kernmenge von Beschreibungsdaten entwickelt. Das sog. *Dublin Core Metadata Element Set* legt eine kleine, überschaubare Menge von Metadaten-Elementen fest. Es können verschiedene Arten digitaler Objekte beschrieben werden, u.a. Töne ('sound'), Bilder ('image') und Texte ('text'). Kategorien zur Beschreibung von Informationsressourcen sind: Titel, Ersteller, Gegenstand, Beschreibung, Beiträger, Verlag, Rechte, Datum[28]. Für die Beschreibung von Korpora und Texten ist diese Metadatenmenge nur bedingt geeignet. Auch aus diesem Grunde hat die *Open Language Archives Community* (kurz: OLAC) die Dublin Core Metadaten um Angaben erweitert, die für Sprachressourcen spezifisch sind[29].
- Der Metadatensatz der *ISLE Metadata Initiative* (IMDI) eignet sich im Prinzip für Sprachressourcen aller Art, wird aber faktisch überwiegend für Korpora gesprochener Sprache und multimodale Korpora verwendet[30].

---

[27] Adresse: `http://dublincore.org/`: „The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards".

[28] Eine sehr knappe, aber nützliche Einführung finden Sie in der Wikipedia: `http://de.wikipedia.org/wiki/Dublin_Core`.

[29] Siehe `http://www.language-archives.org/OLAC/metadata.html`.

[30] IMDI ist unter `http://www.mpi.nl/IMDI/` beschrieben und dokumentiert. Dort finden sich auch einige Sprachressourcen-Projekte, die diese Metadaten verwenden.

- Der *Corpus Encoding Standard* (CES)[31]. Der CES wurde federführend von der *Expert Advisory Group on Language Engineering Standards* (EAGLES) entwickelt. Wie der Name dieses von der EU geförderten Gremiums vermuten lässt, ist dieser Metadatenstandard für Korpora in sprachtechnologischen Projekten entwickelt worden. Dennoch ist der von CES definierte Metadatensatz auch für die Beschreibung linguistischer Korpora geeignet. Dies hängt unter anderem damit zusammen, dass dieser Standard sich an die Konventionen anlehnt, die die *Text Encoding Initiative* (TEI)[32] für ein breiteres Spektrum an Texten und Korpora aufgestellt hat. Die Kategorien des *Corpus Encoding Standard* sind im Großen und Ganzen eine Teilmenge der von der TEI definierten Kategorien, mit einigen wenigen für die Sprachtechnologie relevanten Erweiterungen.
- Die *Component Metadata Infrastructure* (CMDI)[33] versucht, die Proliferation verschiedenster Metadatenstandards für Sprachressourcen dadurch zu überwinden, dass Metadaten, die einen bestimmten Aspekt der Primärdaten beschreiben, als einzelne Komponenten angelegt werden können. Die einzelnen Bestandteile dieser Komponenten wiederum, also die einzelnen Datenelemente, können sich hinsichtlich ihres Namens und ihrer Bedeutung auf einen anderen Metadatenstandard beziehen, nur muss dies explizit gemacht werden. CMDI versteht sich also als ein übergreifendes Informationsgebilde. Daten, die in den anderen Metadatenstandards vorliegen, können in CMDI konvertiert werden. Legt man einen neuen Metadatensatz an, dann schaut man zunächst, ob es im Repositorium von CMDI-Komponenten bereits das für die eigenen Daten passende Beschreibungsschema gibt, ansonsten muss man eine eigene Komponente entwickeln. Der Aufwand, Ressourcen mit CMDI zu beschreiben, ist etwas größer als bei den meisten anderen hier genannten Formaten. Der Gewinn ist, dass sich die Primärdaten leichter mit anderen Ressourcen vernetzen lassen.

Wir werden im Folgenden kurz auf den Aufbau des CES-Metadatensatzes als ein Beispiel eingehen und verweisen im Übrigen auf die oben genannten Webadressen[34]. Metadaten werden im Vorspann eines Textes oder eines ganzen Korpus abgespeichert, sie begleiten also in der Regel die eigentliche Informationsressource. Der Metadatensatz enthält einige wenige Felder, die ausgefüllt werden müssen, und viele Felder, die ausgefüllt werden können.

In einem Feld namens *cesHeader*, welches den Metadatensatz einleitet, kann u.a. erfasst werden, welcher Typ von Informationsobjekt beschrieben wird, wer das Objekt beschrieben hat und welche Version der Metadaten vorliegt. Es handelt sich hierbei also um Meta-Metadaten.

---

31 Siehe `http://www.cs.vassar.edu/CES/CES1.html`.

32 Siehe `http://www.tei-c.org/`: „The Text Encoding Initiative (TEI) Guidelines are an international and interdisciplinary standard that facilitates libraries, museums, publishers, and individual scholars represent a variety of literary and linguistic texts for online research, teaching, and preservation.". Im Zusammenhang der Korpuslinguistik sind vor allem die Kapitel 5 (‚TEI Header') und 23 (‚Language Corpora') von Interesse.

33 Siehe `http://www.clarin.eu/content/component-metadata`.

34 Einen guten und etwas ausführlicheren Überblick über Metadatenstandards für Sprachressourcen gibt das *Benutzerhandbuch* des CLARIN-Projekts, das Sie unter `http://de.clarin.eu/en/language-resources/userguide.html` finden (Text in Englisch).

Der erste Teil des Metadatensatzes (‚file description') beschreibt die bibliographischen Daten des (digitalisierten) Textes oder Korpus. Hierzu gehören u.a. der Titel, die Speichergröße der Datei, Informationen zur Veröffentlichung der Datei sowie Informationen zur Originalquelle, aus der der Text oder das Korpus stammt.

Der zweite Teil enthält Informationen zu Kodierung (‚encoding') der Datei. In diesem Teil wird vor allem das Verhältnis der beschriebenen Informationsressource zum Original beschrieben. Hier können außerdem allgemeine Bearbeitungsrichtlinien angegeben werden.

Der dritte Teil enthält, unter der Bezeichnung *Profil* (‚profile'), eine Reihe zusätzlicher Angaben zum beschriebenen Text. Hierzu gehören u.a. die Textklasse bzw. Textsorte, die Sprache oder Sprachen, in denen der Text verfasst ist, Hinweise auf Übersetzungen des Textes und auf weitere Dateien, in denen auf diesen Text bezogene linguistische Annotationen gespeichert sind.

Im vierten Teil der Metadaten kann schließlich die Revisionsgeschichte der Informationsressource verzeichnet werden, sofern Revisionen an dieser vorgenommen wurden.

Der hier beschriebene *Corpus Encoding Standard* erlaubt eine reichhaltige Beschreibung von Korpora und von einzelnen Texten. Trotz der hohen Zahl an Beschreibungskategorien genügt bereits die Angabe einiger weniger Kategorien, um die Metadaten eines Textes oder Korpus standardkonform zu machen. Der Standard eignet sich so auch für kleinere Projekte und Korpora, bei denen um die Erstellung von Metadaten kein großer Aufwand getrieben werden kann. Seine Anwendung ist also auf jeden Fall eine Überlegung wert.

3.3 Methodische Probleme und ihre Lösung

In den folgenden Abschnitten werden wir auf einige methodische Probleme eingehen, die es beim Aufbau und bei der Verwendung von Korpora zu beachten gilt. Zum Teil trugen diese methodischen Probleme zur Kritik seitens der generativen Linguisten an der Korpuslinguistik bei. Es ist deshalb wichtig, Lösungen für diese Probleme zu entwickeln.

!

3.3.1 Repräsentativität von Korpora

Dieser Abschnitt diskutiert das Verhältnis von Korpora und den Sprachausschnitten, den diese Korpora repräsentieren. Dahinter steckt die Frage, inwieweit man Erkenntnisse, die man durch die Analyse von Korpusdaten gewonnen hat, auf den Sprachausschnitt, den das Korpus repräsentieren soll, übertragen kann. Möchte man zum Beispiel Aspekte der deutschen Sprache der Gegenwart untersuchen, so hat man es bei diesem Untersuchungsobjekt zunächst mit einem nicht präzise abgrenzbaren Phänomenbereich zu tun. Jeden Tag werden Äußerungen in dieser Sprache getätigt, und das meiste entgeht unserer Aufmerksamkeit. In der Terminologie der Statistik spricht man davon, dass die *Grundgesamtheit*, über die man etwas aussagen möchte, nicht präzise definiert werden kann.

Dies ist zum Beispiel bei Meinungsumfragen zum Wahlverhalten der Deutschen anders. Die Grundgesamtheit der wahlberechtigten Deutschen kann hinreichend genau bestimmt werden, um daraus Stichproben zu ziehen, die repräsentativ für die Grundgesamtheit sind. Für die Bestimmung der Stichproben werden Merkmale der Befragten

wie Alter, Herkunft und Bildungsgrad herangezogen, deren Verteilung über die gesamte Bevölkerung ebenfalls bekannt ist. Dadurch lassen sich aus den Ergebnissen der Stichprobenbefragung Schlüsse auf die Gesamtheit ziehen, z.B. über das Wahl- oder Kaufverhalten *der Deutschen*.

In der Korpuslinguistik ist das Verhältnis zwischen Stichprobe und Grundgesamtheit komplizierter. Zwar gibt es Fälle, in denen die Grundgesamtheit abgeschlossen ist, etwa bei den nicht mehr verwendeten und nur schriftlich überlieferten Sprachen, Sprachstufen und Individualsprachen (z.B. das klassische Latein, das Mittelhochdeutsche oder die Sprache Schillers). Bei einer gegenwärtig verwendeten Sprache können wir das Verhältnis von Stichprobe zu Grundgesamtheit nicht exakt bestimmen. Es ist zum Beispiel nicht zu ermitteln, wie groß der Anteil der Fragesätze an allen Sätzen des Deutschen ist. Entsprechend kann dieses Verhältnis nicht in einem Korpus widergespiegelt werden. Dies macht vor allem quantitative Aussagen wie die, dass Modalpartikeln *überwiegend*[35] in Fragesätzen vorkommen, schwer nachprüfbar:

(1) Wurde das *denn / überhaupt* untersucht?

Qualitative Aussagen, etwa dass Modalpartikeln in Fragesätzen eine andere Funktion haben als in Aussagesätzen, sind leichter auf die repräsentierte Gesamtheit übertragbar, stehen aber ebenfalls unter dem Vorbehalt, dass in den untersuchten Daten noch nicht alle Funktionen der Modalpartikeln beobachtet werden konnten[36].

Es gibt mehrere Möglichkeiten, mit dem Problem der Repräsentativität von Korpora und der Verallgemeinerung von Aussagen umzugehen. Wir werden diese im Folgenden vorstellen.

Beschränkung auf das Korpus

Man kann natürlich alle Erkenntnisse, die man durch Beobachtung an einem Korpus gewinnt, allein auf dieses Korpus beziehen. Dies widerspricht aber normalerweise dem Forschungsinteresse der Korpuslinguistik[37]. In der Korpuslinguistik sollen Erkenntnisse gewonnen werden, die über die beobachtete Datenmenge hinaus generalisierbar sind und so unsere Einsicht in die Funktion und Verwendung einer Sprache vertiefen.

Erstellung eines ausgewogenen Korpus

Eine weitere Lösung, die in der korpuslinguistischen Literatur vorgeschlagen wurde, ist, ein ausgewogenes Korpus zu erstellen[38]. Die Ausgewogenheit wird hier vor allem auf

---

[35] In einer konkreten Untersuchung müsste dieser Ausdruck natürlich noch in einen komparativen (z.B. *häufiger als bei allen anderen Satztypen*) oder einen skalaren Term (z.B. in mehr als 60 Prozent der Fälle) überführt werden.

[36] Es wäre z.B. möglich, dass diese Aussage auf Grund der Analyse eines Korpus der gesprochenen Sprache getroffen wurde, die Verhältnisse in der geschriebenen Sprache aber andere sind.

[37] Eine Ausnahme bilden Korpora, die den Gegenstand komplett abdecken, z.B. ein Korpus der Werke Schillers.

[38] Vgl. Atkins et al. (1992). Die Autoren sprechen von einem ‚balanced corpus'. Ausgewogenheit ist auch eines der Kriterien des *Kernkorpus der deutschen Sprache des 20. Jahrhunderts*, vgl. Geyken (2007).

Textsorten bezogen. Der Weg zu einem ausgewogenen Korpus soll durch das Zusammenspiel von externen Kriterien und internen Kriterien erreicht werden.

Zunächst werden Äußerungssorten nach externen Kriterien ausgewählt, z.B. nach der Anzahl der beteiligten Personen (Rede, Interview, Schauspiel etc.) nach dem Grad der Mündlichkeit und Schriftlichkeit (spontanes Gespräch, abgelesene Rede, Chatprotokoll, Zeitungsartikel etc.), nach der Situation (formell, informell etc.)[39]. Diese Kategorien führen zu einer Menge von Textsorten, deren Verteilung in der täglichen Kommunikation beobachtet bzw. geschätzt wird. Diese quantitativen Verhältnisse der Textsorten zueinander werden in einem ersten Schritt des Korpusaufbaus im Korpus widergespiegelt. Es muss allerdings gesagt werden, dass bis heute kein wasserdichtes Verfahren existiert, die Textsorten einer Sprache zu einer gewissen Zeit zu ermitteln. Es wird vermutlich auch nie eines geben. Das Beste, was man erreichen können wird, ist eine pragmatische Lösung, auf die sich die Gemeinschaft beteiligter Korpuslinguisten einigt.

Im Anschluss daran wird bei jeder Äußerungssorte die Verteilung möglichst vieler linguistischer Phänomene beobachtet[40]. Am interessantesten sind dabei diejenigen Phänomene, die durch ihre Häufigkeit und Verteilung für eine Textsorte charakteristisch sind. So ist zum Beispiel die Textsorte *Kochrezept* charakterisiert durch eine hohe Frequenz von Aufforderungssätzen:

(2) Geben Sie nun etwas Zitronensaft in den Teig.

und von befehlssatzähnlichem, subjektlosen Infinitiv-Konstruktionen:

(3) Den Teig fünf Minuten lang gut durchrühren.

Die linguistischen Phänomene, die für eine Äußerungssorte charakteristisch sind, bilden das Profil interner Kriterien für diese Sorte. Man sollte allerdings den Aufwand, der notwendig ist, um linguistische Phänomene in Korpora aufzufinden und quantitativ zu erfassen, nicht unterschätzen. Es ist in jedem Falle schwierig, in manchen Fällen sogar unmöglich, diese Phänomene in einem Korpus automatisch aufzuspüren.

Die Definition interner Kriterien für einzelne Äußerungssorten dient der folgenden Zielen:

- Wenn ein Forscher ein bestimmtes linguistisches Phänomen untersuchen will, oder für seine Untersuchung Daten eines bestimmten linguistischen Phänomens benötigt, dann kann er sich vor allem auf Texte der Sorte stützen, bei der dieses Phänomen häufig vorkommt. Die Auswahl eines solchen Textkorpus wird erleichtert, wenn die internen Kriterien für jede Textsorte im Korpus bzw. dessen Metadaten vermerkt sind[41].

---

[39] Eine ausführliche Liste externer Kategorien findet sich in Atkins et al. (1992).

[40] Wegweisend ist die Arbeit von Biber (1988), der statistische Daten zur Verteilung von mehreren Dutzend linguistischer Eigenschaften in verschiedenen Textsorten – Reportage, Wissenschaftsartikel, schöne Literatur etc. – präsentiert, vgl. Anhang III in Biber (1988), S. 245ff.

[41] So kann man zum Beispiel aus den Korpora des Instituts für deutsche Sprache in Mannheim sog. *virtuelle Korpora* bilden, die aus Texten einer bestimmten Sorte oder mit einem bestimmten Merkmal zusammengestellt werden.

- Der Abgleich des Profils interner linguistischer Merkmale eines Textes mit denen der Textsorten eines Korpus erleichtert die Einordung dieses Textes in das Korpus, falls die Einordnung nicht bereits durch externe Kriterien festgelegt ist.
- Ein Korpus kann auch dadurch ausgewogen gestaltet werden, dass linguistische Phänomene, die generell selten vorkommen, in einem Korpus stärker berücksichtigt werden. Man kann in diesem Fall von einer Austarierung des Korpus nach internen Kriterien sprechen. Das Korpus spiegelt dann nicht mehr unbedingt die Verteilung von Textsorten in der beschrieben Sprache wieder. Es kann aber von Vorteil sein, wenn alle interessanten linguistischen Phänomene in ausreichendem Maße dokumentiert sind. Zudem ist, wie wir gesehen haben, die Repräsentativität einer Stichprobe im Verhältnis zur Grundgesamtheit eine Fiktion, solange die Grundgesamtheit nicht exakt bestimmt werden kann. Es besteht also kein Grund, den Aufbau eines Korpus an einem sowieso nicht genau zu bestimmenden quantitativen Verhältnis zum Gegenstand zu orientieren. Linguistische Kriterien können ebenfalls den Ausschlag geben.

Überprüfung einer Hypothese an mehreren Stichproben

Wie wir oben festgestellt haben, ist ein Korpus immer nur eine Art Stichprobe, von der wir nicht wissen, ob sie wirklich repräsentativ ist und die Verhältnisse so widerspiegelt, wie sie auch in der Gesamtheit sind. Diese Tatsache verhindert aber nicht, dass man linguistische Erkenntnisse über eine Sprache anhand von Korpusdaten gewinnt. Wenn man Hypothesen über linguistische Phänomene auf der Basis von Korpusdaten bildet, muss man sich nur immer im Klaren darüber sein, dass sie eventuell durch die Auswertung einer anderen Stichprobe, also eines anderen Korpus, widerlegt werden könnten. Die Gegenprobe kann entweder vom gleichen Forscher oder von anderen Teilnehmern des korpuslinguistischen Diskurses erbracht werden. Dies entspricht dem normalen Prozess linguistischer Erkenntnis. Zum Beispiel können Erkenntnisse über Frequenz und Verteilung von Modalpartikeln, die anhand eines Korpus der geschriebenen Sprache gewonnen wurden, anhand eines Korpus der gesprochenen Sprache bestätigt oder widerlegt werden. Korpora verschiedener Dialekte oder regionalsprachlichen Varianten des Deutschen können helfen, das Bild von Frequenz und Verteilung der einzelnen Partikel zu verfeinern.

Das Bild, das sich hier ergibt, ist das einer ständigen Verfeinerung der linguistischen Erkenntnisse auf Grund einer immer solideren Materialbasis.

3.3.2 Prognose vs. Korpusevidenz

Eine wichtige Aufgabe der modernen Korpuslinguistik ist es, die intuitiven Entscheidungen und Theorien von Linguisten an großen Mengen authentischer Sprachdaten zu überprüfen. Dabei wird die Intuition der Linguisten und befragten Sprecher bestätigt oder korrigiert werden. Man spricht davon, dass bestimmte Aussagen, die von Linguisten auf Grund einer bestimmten Theorie getroffen werden, anhand von Korpusevidenz verifiziert werden. Die Frage ist allerdings, wie stark diese Evidenz sein muss, damit sie als Gegenpol zu theoretischen Aussagen anerkannt werden kann. Wir wollen hier in Erinnerung rufen, dass in Korpora nicht nur Sätze vorkommen, die wohlgeformt sind. Wir haben es mit einer nicht zu unterschätzenden Zahl von Äußerungen zu tun, die

ungrammatisch sind oder deren Grammatikalität zumindest zweifelhaft scheint. Andererseits werden in linguistischen Arbeiten und auch in Grammatiken Sätze als Beispiele herangezogen, deren Verwendung in authentischen Äußerungen äußerst unwahrscheinlich ist. Gill Francis zitiert das folgende Beispiel aus einer Grammatik von Quirk und anderen[42]:

(4) Walter played the piano more often in Chicago than his brother conducted concerts in the rest of the states.

Francis bezeichnet diese Art von Beispielsätzen als grammatisch, aber unnatürlich und ihre Verwendung als höchst unwahrscheinlich[43].

Es wird also eine Vielzahl von Konstruktionen geben, die zwar bildbar und grammatisch sind, die man aber mit hoher Wahrscheinlichkeit in keinem Korpus finden wird. Zur Überprüfung der Wohlgeformtheit solcher Konstruktionen bleibt deshalb nur die Befragung von Muttersprachlern.

Wie geht man aber mit der Situation um, dass im Korpus Belege für Konstruktionen gefunden werden, die im Kontext einer Theorie oder Grammatik als nicht wohlgeformt eingestuft werden? Diese Frage ist schwerer zu beantworten[44]. Wir wollen an zwei Beispielen zeigen, wie man mit dieser Situation umgehen kann.

Das erste Beispiel stammt von Detmar Meurers[45]. Meurers verwendet Korpusdaten, um eine Hypothese von den Besten und Edmondson zu überprüfen. Diese behaupten[46], dass Sprecher einiger süddeutscher Dialekte eine sonst nicht vorkommende Anordnung von Verben innerhalb einer komplexen Verbalgruppe verwenden, wie im folgenden Beispiel:

(5) dass er singen hat müssen

Das Besondere an diesem und ähnlichen Beispielen ist die Stellung des finiten Verbs (hier *hat*) zwischen zwei von ihm abhängigen infiniten Verben. Den Besten und Edmondson erklären die Verwendung dieser Konstruktion als das Bemühen der Dialektsprecher, hochsprachlich zu klingen, also als eine Art Überkompensation, wenn man davon ausgeht, dass diese Konstruktion ungrammatisch ist[47]. Den Befunden von den Besten und Edmondson folgend, müsste eine Grammatik des Deutschen diese Konstruktion entweder ausschließen, denn es handelt sich um ein reines Produkt der Performanz, um eben den Versuch der Anpassung an eine nicht existente Norm. Alternativ könnte diese Konstruktion in eine regional ausdifferenzierte Grammatik des Deutschen als Besonderheit der bairischen Dialekte aufgenommen werden.

---

42 Vgl. Francis (1993), S. 139. Francis bezieht sich hier auf die *Comprehensive Grammar of the English Language*, erschienen 1985.

43 Ebd.

44 Eine neue Antwort auf diese Frage formulieren András Kertész and Csilla Rákosi mit dem Konzept der Plausibilität einer linguistischen Aussage einer Theorie, s. dazu Abschnitt 2.2 in diesem Buch.

45 Vgl. Meurers (2005), Kapitel 1.3.

46 Vgl. den Besten und Edmondson (1983), S. 182.

47 Sie können Ihre eigene Intuition in dieser Frage prüfen, indem Sie den obigen Satz mit den folgenden Varianten vergleichen: a) *dass er hat singen müssen*, b) *dass er singen gemusst hat*.

Meurers durchsucht ein Zeitungskorpus[48] nach Beispielen für die Konstruktion in Beispiel (5) und wird fündig. Er findet insgesamt zehn Belege, die dieser Konstruktion entsprechen und mutmaßt, dass es angesichts dieses Befundes sinnvoll sein könnte, diese Konstruktion als grammatisch zu markieren[49].

Die Korpusbefunde, die Meurers präsentiert, sind nicht wirklich überzeugend als Gegenargument zu den Besten und Edmondsons Argumenten. Es könnte tatsächlich sein, dass die Beispielsätze von Sprechern des Bairischen verfasst wurden. Damit wären sie als Teil eines regionalen Sprachgebrauchs bzw. als Phänomen der Performanz deutbar. Über die Herkunft der Verfasser dieser Sätze wissen wir leider nichts. Die methodische Frage lautet: Wie viele Schwalben machen einen Sommer? Auf unser Problem übertragen: Wie viele Belege deuten auf eine Regularität hin, die wir bei einer linguistischen Beschreibung berücksichtigen müssen? Dies ist eine sehr interessante und offene Forschungsfrage, mit der sich die Korpuslinguistik unseres Erachtens bisher zu wenig befasst hat.

Anders geht Geoffrey Sampson in seiner Auseinandersetzung mit theoretischen Linguisten um das Phänomen des *central embedding* vor[50]. Sampson widerlegt mit der Hilfe von Korpusbelegen die Behauptung von theoretischen Linguisten, dass die mehrfache Einbettung – X erscheint eingebettet in X, welches wiederum in X eingebettet ist, usw. – kein natürlich auftretendes Phänomen ist. Diese Strukturen seien danach zwar grammatisch, aber unakzeptabel. Sampson, einmal auf dieses Phänomen aufmerksam geworden, sammelt aus verschiedenen Quellen authentischen Sprachgebrauchs eine Vielzahl von Belegen für diese Struktur. Insgesamt fünfzehn davon präsentiert er in seinem Aufsatz. Wenn man seine Vorgehensweise systematisieren würde, dann erhielte man die Methode „Überprüfung einer Hypothese an mehreren Stichproben", wobei die Hypothese hier lautet: Konstruktionen dieser Art werden verwendet.

Zusammenfassung: Wenn Sie eine grammatische Konstruktion, deren Korrektheit aus Ihrer Sprachtheorie folgt, anhand von Korpusdaten überprüfen wollen, dann kann es sein, dass Sie diese in Ihren Korpora nicht finden. In diesem Fall bleiben Ihnen andere Möglichkeiten der Bestätigung, z.B. indem Sie Muttersprachler befragen. Zweifeln Sie andererseits eine Konstruktion, deren Korrektheit aus einer bestimmten Sprachtheorie folgt, an, dann ist die Tatsache, dass es keine Korpusbelege für sie gibt, noch kein hinreichendes Argument. Auch hier kann möglicherweise die Befragung von Muttersprachlern entscheiden.

Wenn Sie jedoch zeigen wollen, dass eine grammatische Konstruktion verwendet wird, die nach Auffassung einer Sprachtheorie nicht wohlgeformt bzw. ungrammatisch

---

[48] Das Korpus besteht aus Texten der Frankfurter Rundschau. Es umfasst etwa 2,6 Millionen Sätze bzw. gut 35 Millionen Wörter. Wichtig ist es, dass das Korpus eine Stichprobe des hochdeutschen, nicht des bairischen, Sprachgebrauchs ist.

[49] Meurers schreibt: „One is bound to ask how such verbal complex patterns could be licensed for those speakers who find them grammatical." Die Formulierung ist äußerst vorsichtig, es bleibt aber zu fragen, ob der Vorschlag nicht dazu führen würde, für jeden Sprecher eine eigene Grammatik, entsprechend seiner Intuitionen, zu entwickeln.

[50] Vgl. Sampson (1996): „Central embedding refers to structures in which a constituent occurs medially within a larger instance of the same kind of ragma; an invented example is [*The book* [*the man left*] *is on the table*], where a relative clause occurs medially within a main clause . . .", S. 15. Wir stützen uns bei der folgenden Darstellung auf diesen Text.

ist, dann ist die Argumentation schwieriger. Es gibt bisher keine theoretisch ausreichend fundierte Methode, um korrekte von nicht korrekter Sprachverwendung zu unterscheiden. Die Belege, die Sie präsentieren, können deshalb immer als nicht korrekter Sprachgebrauch disqualifiziert werden. Man kann beim jetzigen Stand der Korpuslinguistik nur pragmatisch vorgehen. Je mehr Belege für die zweifelhafte Konstruktion gefunden werden, und je vielfältiger die Fundstellen sind, um so gesicherter kann man die Existenz dieser Konstruktion behaupten und darauf bestehen, dass die Theorie den beobachteten Fakten angepasst wird.

3.4 Methodisches Vorgehen beim Aufbau eines Korpus — Eine Anleitung

Am Schluss dieses Kapitels wollen wir für den Fall, dass Sie ein eigenes Korpus aufbauen wollen, einige Tipps geben:

- Die erste Frage dürfte sein, wie Sie an die Daten herankommen. Da heute praktisch alle Texte bereits in der Druckvorstufe digitalisiert sind, dürfte das Scannen oder die manuelle Eingabe nur noch eine geringe Rolle spielen. Bei älteren Texten werden Sie aber nicht darum herum kommen. Der Aufwand für diese Aufgabe sollte nicht unterschätzt werden. Eine gute Adresse für Texte aller Art ist das World Wide Web. Aber auch, wenn Sie von dort Daten sammeln, müssen Sie einigen Aufwand für die Bereinigung dieser Daten einplanen. Es gibt aber Werkzeuge, die diese Aufgabe unterstützen[51].
- Sie sollten sich so früh wie möglich Gedanken über das Urheberrecht an den von Ihnen gesammelten Daten machen. Ein interessanter, auch für den juristischen Laien zugänglicher Artikel ist Lehmberg et al. (2008). Im Rahmen des deutschen CLARIN-Projektes wird ein „Legal Helpdesk"[52] aufgebaut. Dort findet sich eine Liste mit meist online zugänglicher Literatur zu diesem Thema[53], die Sie konsultieren können. Am besten ist es, wenn Sie mit den Rechteinhabern frühzeitig in Kontakt treten und Ihre Nutzung der Texte durch eine Lizenz rechtlich absichern. Dies dürfte nicht so schwer sein, wenn Sie die Daten ausschließlich zu Forschungszwecken nutzen. Etwas schwieriger dürfte es werden, wenn Sie die Daten weitergeben wollen. Es ist einerseits sinnvoll oder sogar notwendig, dass andere Forscher das selbe Korpus verwenden können, und sei es nur, um ihre Ergebnisse nachprüfen zu können. Andererseits kann dies die Vereinbarung über Nutzungsrechte erschweren[54].

---

[51] Auf der Webseite, die dieses Buch begleitet, stellen wir einige dieser Werkzeuge vor.

[52] Siehe `http://www.clarin-d.de/de/schulungen-und-support/rechtliche-fragestellungen`. Wir danken Erik Ketzan vom Institut für Deutsche Sprache für den Hinweis.

[53] Siehe `http://www.clarin-d.de/de/legal_issues-bibleography`.

[54] Ein Extremfall ist sicher die Arbeit von Christa Dern (2003), die für ihre Untersuchung ein Korpus von Erpresserbriefen verwendete. Es liegt in der Natur der Sache, dass viele Autoren sich nicht ausfindig machen lassen und das Bundeskriminalamt (BKA) als sekundärer Rechteinhaber nur bedingt Interesse an der Verbreitung dieses Korpus hat. Das macht es schwierig, die von Dern präsentierten Erkenntnisse zu Formen der Höflichkeit in Erpresserbriefen und die Schlussfolgerungen der Autorin zu überprüfen. Inzwischen (Stand 2015) gibt das BKA das Korpus auf Anfrage für konkrete Forschungsvorhaben weiter.

- Ein nicht unwesentlicher Aspekt ist die Kodierung der Daten. Moderne Betriebssysteme verwenden heute UNICODE[55], eine Kodierung, mit der sich Zeichen aller Sprachen darstellen lassen. Es sind aber auch noch verschiedene Formate eines von der *International Standardisation Organisation* normierten Zeichensatzes in Gebrauch (z.B. ISO-8859-1 – ISO-8859-15), ebenso wie der wesentlich ältere Kodierungsstandard ASCII ('American Standard Code for Information Interchange'). Man sollte sich über die Kodierung der Textdateien frühzeitig informieren und für alle Dateien die gleiche Kodierung wählen, was eventuell die Konvertierung einiger Dateien erforderlich macht. Der umfassendste Standard ist UNICODE, wir wollen dessen Verwendung deshalb an dieser Stelle empfehlen.
- Spätestens wenn die Primärdaten gesammelt sind, stellt sich die Frage nach den Metadaten. Wir haben oben beschrieben, wann die Beschreibung der Primärdaten durch Metadaten sinnvoll ist: Wenn mehrere Forscher die Daten verwenden und wenn die Daten in einer Forschungsarbeit dokumentiert werden müssen.
- Je nach Forschungszweck kann es sinnvoll sein, die Daten linguistisch zu annotieren. Wir werden in den folgenden Kapiteln ausführlich auf diesen Aspekt der Korpusaufbereitung eingehen.

3.5 Weiterführende Literatur

Für die in diesem Kapitel angeschnittenen Themen ist das Buch von Tony McEnery, Richard Xiao und Yukio Tono (2006) eine ausgezeichnete Referenz, besonders Teil A. Viele der Themen werden auch in einem Aufsatz von Atkins, Clear und Ostler (1992) behandelt. Jeremy Clear geht an anderer Stelle auf die Frage der Repräsentativität und des Aufbaus von Korpora unter diesem Gesichtspunkt ein (1992). Die Frage, ob und wie das World Wide Web als Korpus für linguistische Untersuchungen verwendet werden kann, ist hochaktuell. Eine gute Einführung in die Thematik geben Kilgarriff und Grefenstette (2003). Es gibt außerdem zu diesem Thema eine jährliche Konferenz. Details dazu lassen sich über eine Suchmaschine (z.B. mit dem Stichwort *WaCky*[56]) ermitteln. Der CES-Metadatenstandard ist auf der CES-Webseite (`http://www.cs.vassar.edu/CES/`) sehr gut dokumentiert. Lesenswert, wenn auch leider nur auf Englisch verfügbar, ist der in das Thema Metadaten einführende Text von Lou Burnard, einem der führenden britischen Korpusexperten (`http://ota.ox.ac.uk/documents/creating/dlc/chapter3.htm`). Auf die hier nur angerissenen Themen der linguistischen Annotation und der Korpusabfrage gehen wir in den Kapiteln 4 und 5 näher ein.

---

[55] Details finden Sie unter `www.unicode.org`. Den meisten Lesern dürfte dieser Standard unter dem Namen „utf-8" geläufig sein; streng genommen handelt es sich dabei um eine für die Zeichensätze westeuropäischer Sprachen besonders effiziente Art der Kodierung von UNICODE-Zeichen.

[56] *WaCky* steht für 'The Web-As-Corpus Kool Yinitiative'.

3.6 Aufgaben

1. Nennen Sie jeweils mindestens eine Aufgabe für die sich a) ein komplettes Korpus, b) eine Belegsammlung als Datenbasis gut eignet.
2. Sie wollen aus Texten, die ein Programm für Sie aus dem World Wide Web herunterlädt, ein Korpus der deutschen Sprache aufbauen. Welche Möglichkeiten haben Sie, um möglichst sicherzugehen, dass nur deutschsprachige Texte in Ihrem Korpus landen. Das Korpus wird am Ende zu groß sein, als dass Sie jeden Text einzeln daraufhin überprüfen könnten.
3. Erstellen Sie für das Buch, dass Sie gerade lesen, einen Metadatensatz a) nach dem Dublin Core Modell, b) nach dem CES Modell. Gibt es Informationen, die Sie gern in die Schemata eingetragen hätten, die Sie aber nicht ermitteln konnten?

4 Auf den Schultern anderer stehen — Linguistische Annotationsebenen

Warum alles selber machen? Manche Aufgaben der linguistischen Analyse und Beschreibung sind bereits von anderen erledigt worden. Die für das Deutsche verfügbaren „Produkte" sollen hier beschrieben werden. Es gilt: Man muss essen, was und wie es auf den Tisch kommt. Am Ende dieses Kapitels haben Sie Annotationen auf verschiedenen linguistischen Ebenen kennengelernt: Wortarten, weiterführende syntaktische Kategorien und Relationen, semantische Markierungen, pragmatische Koreferenz- und Diskursrelationen sowie eine Ebene der Normalisierung. Sie sind mit dem sehr beliebten Stuttgart-Tübingen-Tagset (STTS) zur Annotation von Wortarten vertraut. Im Bereich der weiterführenden syntaktischen Annotation können Sie den konstituentenbasierten vom dependenzbasierten Ansatz unterscheiden.

4.1 Motivation

In Abschnitt 2.4 wurde erwähnt, dass Vertreter des Kontextualismus bzw. des korpusbasierten, quantitativ-qualitativen Ansatzes mit rohen Korpusdaten arbeiten. Sie leiten ihre Analysen aus der Datengesamtheit ohne oder nur mit minimalen linguistischen Voraannahmen ab. In diesem Kapitel nehmen wir einen gegensätzlichen Standpunkt ein und argumentieren für den Nutzen, annotierte Daten auszuwerten bzw. die Rohdaten eines Korpus systematisch mit linguistischen Analysen in der Form von Annotationen anzureichern. Die Idee ist, dass Annotationen als eine Art Anker dienen können, und dem Nutzer ermöglichen, auf effiziente Weise relevante Beispiele in einem Korpus zu finden. Annotationen können als kontextualisierte Analysen betrachtet werden. Sie machen Untersuchungsergebnisse für andere Forscher nachvollziehbar und auch überprüfbar, weil die Datengrundlage offengelegt wird. Außerdem stellen Annotationen meistens Generalisierungen über die einzelnen Wortformen dar, und können dabei helfen, interessante linguistische Muster in den Daten zu erkennen. Das Entwickeln eines Annotationsschemas und auch der Annotationsprozess selbst unterstützen den Analyseprozess, da die zugrunde gelegten Konzepte und Definitionen im Abgleich mit den Daten immer wieder auf den Prüfstand kommen[1]. Im Folgenden gehen wir auf drei Aspekte der Motivation von Annotation etwas genauer ein[2].

---

[1] Siehe auch Abschnitt 5.3.

[2] Die folgende Darstellung orientiert sich an Leech (1997).

4.1.1 Extraktion von linguistischer Information

Jeder, der schon einmal in einer Linguistikklausur über Satzanalysen geschwitzt hat, weiß, dass der reine Text nur wenig linguistische Information an der Oberfläche offenbart. Linguistische Kategorien wie die Wortart *Artikel* oder die syntaktische Funktion *Subjekt* lassen sich nicht unmittelbar vom Text ablesen, sondern verlangen, dass man den Text linguistisch interpretiert. Zur Abstraktheit von linguistischen Konzepten kommt erschwerend hinzu, dass das, was man da interpretieren möchte, für sich genommen oft mehrdeutig ist, also mehr als eine Interpretation erlaubt. Diese Mehrdeutigkeit verschwindet allerdings meistens, wenn man den Kontext einbezieht. Die konkrete Äußerungssituation – auch im Sinne von geschriebenem Text – ist meistens ausreichend, um eine mehrdeutige Form zu disambiguieren.

Die Wortform *einen* z.B. hat mindestens drei Lesarten, die auf drei verschiedene Wortarten zurückgeführt werden können. Bevor Sie weiterlesen, überlegen Sie kurz, um welche Lesarten es sich hier handeln könnte.

Die Beispiele (1) – (3) illustrieren die drei Lesarten (Hervorhebungen durch uns)[3].

(1) *Indefiniter Artikel*
Diese Perspektive ermögliche *einen* neuen Blick auf gesellschaftliche Verhältnisse.

(2) *Indefinitpronomen*
Gleichzeitig lautet der Appell an die Mieter, sich doch *einen* der Tiefgaragenplätze anzumieten.

(3) *Verb*
[Sie] wollten [...] von Bremen aus die Republik wieder *einen*.

Haben wir die einzelnen Vorkommnisse erst einmal interpretiert, können wir die Analysen bzw. Annotationen als Grundlage für eine weiterführende linguistische Analyse heranziehen. In Beispiel (1) können wir z.B. die Sequenz *einen neuen Blick auf gesellschaftliche Verhältnisse* zu einer Nominalphrase mit dem Kern *Blick* zusammenfassen, die in Bezug auf das Verb *ermögliche* die Funktion des Akkusativobjekts einnimmt.

Die drei Lesarten von *einen* sind in durchschnittlichen Korpora nicht gleichverteilt. Ist man an der Artikellesart von *einen* interessiert, erhält man bei einer Suche auf den Primärdaten, d.h. den reinen Wortformen, viele relevante Beispiele. Anders sieht es aus, wenn man an der viel selteneren Verblesart, vgl. (3), interessiert ist. In diesem Fall müsste man voraussichtlich eine große Anzahl von irrelevanten Treffern sichten, um auf einschlägige Belege zu stoßen. Diese mühsame und zeitaufwändige Arbeit wird vereinfacht, wenn das Korpus mit Wortartenannotationen angereichert ist. In diesem Fall kann man gezielt nach Kombinationen von *einen* mit einer verbalen Annotation suchen und so die Treffersichtung auf tatsächliche Verbvorkommen einschränken.

---

[3] Es handelt sich hierbei, wie bei vielen Beispielen in diesem Kapitel, um ggf. leicht gekürzte Korpusbelege aus der *Tübinger Baumbank des Deutschen/Zeitungskorpus* (kurz *TüBa-D/Z*). Das *Z* in *TüBa-D/Z* unterscheidet diese Baumbank von der verwandten *TüBa-D/S*: Tübinger Baumbank des Deutschen/Spontansprache. Wir halten es allerdings mit Pullum (2003) und verwenden auch eigens konstruierte Beispiele, falls es der besseren Veranschaulichung dient.

Ein analoges Argument gilt auch für die Suche nach weiterführenden linguistischen Phänomenen im Korpus: Zum Beispiel eine Suche nach möglichen Objekten des Verbs *einen*, wie die Nominalphrase *die Republik* in Beispiel (3), kann wesentlich effizienter durchgeführt werden, wenn Wortgruppen mit syntaktische Phrasen oder syntaktischen Funktionen annotiert sind.

Noch deutlicher wird die Sinnhaftigkeit von Annotation, wenn man ein linguistisches Phänomen untersucht, für das man noch keine einschlägigen Wortformen benennen kann, sondern diese erst aus dem Korpus ermitteln möchte. Ein Beispiel hierfür ist die Untersuchung von Prädikativkonstruktionen im Genitiv[4]. Eine Suche auf der syntaktisch annotierten TüBa-D/Z liefert u.a. folgende Treffer: *der Ansicht sein*, *der Meinung sein*, *guten Mutes sein*. Die weitere Interpretation der Ergebnisse, in wie weit es sich hier tatsächlich um Prädikativkonstruktionen gemäß einer bestimmten Theorie handelt, liegt dann in der Hand des Linguisten oder Lexikographen.

Sprache kann in vieler Hinsicht mehrdeutig sein – nicht nur auf der Wortebene. Eine strukturelle Ambiguität kann z.B. beim Bezug von Präpositionalphrasen bestehen (beim sogenannten *PP-Attachment*). Unter Linguisten ist in diesem Zusammenhang ein Zitat von Groucho Marx berühmt: „Last night I shot an elephant in my pajamas and how he got in my pajamas I'll never know"[5]. Diese Ambiguität des Bezugs von *in my pajamas* ist für den Leser eine Falle, da sie im Folgesatz in die weniger wahrscheinliche Lesart aufgelöst wird. Im Korpus kann sie durch syntaktische Annotation eindeutig festgehalten werden, indem *in my pajamas* als Attribut der nominalen Struktur von *an elephant* zugeordnet wird – und nicht als Umstandsangabe dem verbalen *shot*.

Die Beispiele haben gezeigt, dass es sinnvoll ist, Korpusdaten mit linguistischen Interpretationen anzureichern, indem man z.B. Wortarten, syntaktische Phrasen oder grammatische Funktionen annotiert. Diese Annotationen machen Korpusabfragen effizienter, indem präzisere Anfragen gestellt werden können und abstrakte Konzepte in den Daten abfragbar gemacht werden.

4.1.2 Wiederverwendbarkeit

Die oben beschriebene Interpretation von Daten ist zeitaufwändig. Viel Zeit geht verloren, wenn jeder dieselben Texte immer wieder neu interpretieren muss. Ein annotiertes Korpus ist auch deshalb wertvoll, weil es erlaubt, die Interpretationen anderer zeitsparend zu nutzen.

Ein zweiter Aspekt der Wiederverwendbarkeit bezieht sich auf die Korpusannotation selbst. In vielen Korpusprojekten werden Programme zur automatischen Annotation von Wortarten verwendet (*Wortarten-Tagger*). Die automatische Bestimmung einer Wortart ist oft nur dann möglich, wenn die Wortarten der unmittelbar umgebenden Wörter ebenfalls bestimmt werden: Nach einem Artikel wie *eine* ist das Wort *lange* mit hoher Wahrscheinlichkeit ein Adjektiv wie in *eine lange Pause*. Steht *lange* jedoch unmittelbar vor einem Partizip, handelt es sich eher um ein Adverb wie in *Sie hat lange gewartet*.

---

[4] Dieses Suchbeispiel geht auf eine Anfrage von Judith Berman zurück.

[5] Siehe `http://groucho-marx.com`.

Die verschiedenen Ebenen der Interpretation bauen oft aufeinander auf. Auch hier ist die Wiederverwendbarkeit von bereits erarbeitetem Wissen wertvoll. Eine Sequenz von drei Wörtern, die mit den Wortarten *Artikel Adjektiv Nomen* annotiert ist (wie *eine lange Pause*), kann auf der Ebene der syntaktischen Annotation automatisch zu einer Nominalphrase zusammengefasst werden.

4.1.3 Multifunktionalität

Ein weiterführender Aspekt von Wiederverwendbarkeit ist der Einsatz derselben Ressource in ganz unterschiedlichen Bereichen. Zum Beispiel kann ein Korpus zur Erstellung oder Verbesserung eines Lexikons erstellt worden sein. Die Lexikographen waren eventuell an Informationen über mögliche Valenzrahmen von Verben interessiert oder an Kollokationen. Dasselbe Korpus kann dann auch von Computerlinguisten genutzt werden, um einen syntaktischen Parser oder andere computerlinguistische Werkzeuge zu entwickeln[6].

Stellen Sie sich vor, Sie würden selbst ein Korpus zur Kommunikation in Chaträumen erstellen, weil Sie an der Verwendung von Neologismen und Anglizismen in dieser informellen Sprachvariante interessiert sind. Es wäre dann gut denkbar, dass andere auf Sie zukommen, um Ihr Korpus für ganze andere Fragestellungen zu nutzen z.B., um die Verwendung von Modal- und Abtönungspartikeln in dieser konzeptuell mündlichen Varietät zu studieren.

4.2 Grundlagen

4.2.1 Übersicht zum Annotieren und zu Annotationsebenen

Nachdem der Einsatz von linguistischer Annotation begründet wurde, geht es in den folgenden Abschnitten um die Sache selbst. Welche Arten von Annotationen findet man in Korpora? Hierzu ist noch eine kurze Vorbemerkung nötig und zwar zur Frage, wie Annotationen erstellt werden.

Wie wir oben schon betont haben, ist das Erstellen von Annotationen zeitaufwändig und dadurch teuer. Beim Annotieren wird daher oft zweistufig vorgegangen: Zuerst findet ein schneller, *automatischer* Vorverarbeitungsschritt statt, bei dem ein computerlinguistisches Werkzeug (auch *Tool*) zum Einsatz kommt. Mithilfe von Regeln oder aus Korpora abgeleiteten Wahrscheinlichkeiten reichert es die Textdaten mit Annotationen an. Im zweiten Schritt ergänzen oder korrigieren Annotatoren – oftmals studentische Hilfskräfte – die automatische Annotation in einem *manuellen* Arbeitsschritt. Es gibt auch interaktive Annotationsprogramme, bei denen die strikte Teilung der Arbeitsschritte aufgehoben ist[7]. Das Programm schlägt dem Annotator aufeinander aufbauende Teilanalysen vor, die unmittelbar korrigiert werden können, so dass die einzelnen Reparaturschritte klein bleiben und effizient durchführbar sind.

Automatische Annotationstools sind von unterschiedlicher Qualität und allesamt nicht perfekt. Normalerweise geht man einen Kompromiss ein zwischen der Größe der

---

[6] Siehe auch Abschnitt 8.6.

[7] Zum Beispiel das Programm *Annotate* (Brants und Plaehn, 2000) für syntaktische Annotation.

annotierten Datenmenge auf der einen Seite und der Qualität der Annotation auf der anderen. Man muss hier allerdings darauf hinweisen, dass auch manuelle Annotation nicht automatisch mit einer fehlerfreien Annotation gleichzusetzen ist. Wenn mehrere Annotatoren die gleichen Daten annotieren, stimmen sie selten hundertprozentig überein[8]. Man versucht, die Abweichungen möglichst gering zu halten, indem man explizite Annotationsrichtlinien (*Annotation Guidelines*) mit Definitionen und Beispielen für alle Annotationskategorien vorgibt und Entscheidungshilfen für problematische Fälle bereitstellt. Die Annotationskategorien werden als *Tags* bezeichnet. Ein Tag (gesprochen [tæg]) kann einem einzelnen Wort, einer Sequenz oder auch einer Relation zugeordnet werden. Im Falle von Wortartenannotation z.B. benennen die Tags die einzelnen Wortartenklassen wie *Artikel* oder *Präposition*.

Tabelle 2 gibt eine kleine Übersicht zu gängigen linguistischen (und computerlinguistischen) Annotationsebenen.

Ebene	Annotation
Morphosyntax	Wortart (*Part of Speech*)
Morphologie	Grundform (*Lemma*), Flexionsmorphologie
Syntax	Konstituenten oder Dependenzen, oft mit syntaktischen Funktionen; andere strukturelle Organisationsform: Topologische Felder
Semantik	Eigennamenklassen, Lesarten (*Word Senses*), thematische Rahmen (*Frames*), Zeitausdrücke und Bezüge
Pragmatik	Koreferenz, Informationsstatus, Informationsstruktur, Diskursrelationen, Konnotation (*Sentiment*)
Weitere	Textstruktur, Orthographie, Normalisierungsebenen, Fehlerannotation, phonetische und prosodische Merkmale, Pausen, sprachbegleitende Merkmale wie Gestik und Mimik und vieles mehr

Tabelle 2: Gängige linguistische Annotationsebenen

4.2.2 Segmentierung

Das Thema dieses Abschnitts mag zunächst etwas überraschen. *Segmentierung* bedeutet schließlich Aufteilung und nicht Hinzufügung, wie man es bei Annotation erwarten würde. Um die Annotation in Korpora nachvollziehen zu können, muss man sich zunächst über die Einheiten im Klaren sein, die man mit einer Annotation markieren möchte. Ein Text muss dazu in seine Bestandteile zerlegt werden. Die Segmentierung kann bei der *Textstruktur* beginnen und Bestandteile eines Textes wie Kapitel,

[8] Als Wert für die Übereinstimmung (*Inter-Annotator Agreement*) wird neben dem prozentualen Anteil der übereinstimmenden Annotationen oft das sog. κ-Maß (kappa) angegeben, welches berücksichtigt, dass ein Teil der Übereinstimmungen dem Zufall geschuldet ist, vgl. Artstein und Poesio (2008); Perkuhn et al. (2012).

Überschrift, Vorspann, Grundtext, Bildtext, Fußnote, Paragraph usw. markieren. Sie geht bis zum *Satz* und unterteilt diesen wiederum in einzelne Worteinheiten. Diese beiden letzten Zerlegungsschritte werden auch unter dem Schlagwort *Tokenisierung* zusammengefasst[9].

Man könnte meinen, dass die Festlegung von Satzgrenzen keine Schwierigkeiten bereite. Für die automatische Erkennung von Satzgrenzen stellt die Disambiguierung des Punktzeichens jedoch eine echte Herausforderung dar, die über Regeln und Statistiken gelöst werden muss. Beispiel (4) illustriert drei Lesarten des Punktes: Abkürzungspunkt, Ordinalzahlenpunkt und Satzendepunkt. Können Sie die drei Lesarten des Punktes im Beispiel identifizieren?

(4) Prof. Dr. Marga Reis eröffnete die Konferenz am 2. Februar mit einem Grußwort.

Beispiel (5) zeigt, dass der Punkt, der auf eine Zahl folgt, nicht immer eine Ordinalzahl markiert.

(5) Es begann 2002.

Die weitere Zerlegung in Worteinheiten identifiziert nicht nur Wörter im gängigen Sinn als *Token*, sondern auch Zahlen, Satzzeichen, Klammern, Anführungsstriche und andere Symbole. Die einfachste Methode dabei ist, sich an Leerstellen zu orientieren und anzunehmen, dass eine geschlossene Zeichenfolge zwischen zwei Leerstellen eine Worteinheit darstellt. Dass es auch in diesem Bereich Diskussionsbedarf gibt, illustrieren die folgenden Beispiele.

Wie behandelt man *kontrahierte Formen* wie die Verschmelzung einer Präposition mit dem definiten Artikel zum Beispiel bei *am* oder *ins*. Soll *machen's* ein Token sein oder zwei? Was ist mit *glaubense* (= *glauben Sie*)? Und wie soll man mit Wörtern umgehen, die unabsichtlich zusammengeschrieben wurden wie *einKooperationsabkommen*? Ist man dem Originaltext treu samt seiner Formatierung oder korrigiert man den Fehler im Korpus?

Das umgekehrte Problem entsteht bei *Mehrwortlexemen* wie *en bloc* oder *New York*, d.h. Sequenzen, die Leerstellen enthalten, aber gemeinhin als eine Worteinheit empfunden werden. Soll man sie getrennt oder als Einheit betrachten? Wie viele Token umfasst z.B. die Sequenz *1 1/2 Stellen* – zwei, drei oder sogar fünf? Noch schwieriger wird es, wenn Namen oder Idiomatik ins Spiel kommt. Wir überlassen es Ihnen, sich zu überlegen, nach welchen Regeln Sie im folgenden Beispiel die Wortgrenzen festlegen würden.

(6) des „Für alle Fälle Fitz"-Teams

Das Beispiel steht exemplarisch für alle Titel und Bezeichnungen, die intern eine phrasale Struktur aufweisen, im äußeren Satzzusammenhang aber wie eine nicht weiter zer-

[9] Die Zerlegung muss nicht auf der Wortebene aufhören. Bettina Zeisler und Andreas Wagner (2004) beschreiben z.B. die Segmentierung auf Morphemebene für ein Korpus des Tibetischen. Bei Korpora, die Transkriptionen von mittelalterlichen Handschriften enthalten, ist es z.B. sinnvoll, zusätzlich auf der Zeichenebene zu trennen, um Initialbuchstaben oder Farbinformationen annotieren zu können, vgl. Lüdeling et al. (2005a).

legbare Einheit fungieren. Analysiert man sie als einzelne Token, erhält man in späteren Analyseschritten eventuell seltsame Teilstrukturen, weil sie im größeren Zusammenhang nicht der normalen Wortabfolge oder Syntax konform gebildet sind.

4.3 Annotationsebenen im Detail

4.3.1 Morphosyntaktische Annotation

Am meisten verbreitet ist die Annotation von morphosyntaktischer Information. Vereinfacht gesagt handelt es sich hierbei um die Zuweisung der Wortart zu einzelnen Token. Im Englischen heißt die Annotation morphosyntaktischer Merkmale auch *Grammatical Tagging*, *Part-of-Speech Tagging* (kurz: *POS Tagging*)[10] oder einfach *Tagging*[11]. Das Wortartentag erlaubt die Disambiguierung mehrdeutiger Wortformen (Homographen), insofern sie verschiedenen Wortarten angehören. Die Liste aller verwendeten Wortartentags wird als *Tagset* bezeichnet. Wenn man als Linguist bei dem Stichwort *Tagset* eine überschaubare Liste wie Nomen (Substantiv), Verb, Präposition, Konjunktion usw. erwartet, ist man wahrscheinlich überrascht, wenn man die große Anzahl an unterschiedlichen Tags in einem annotierten Korpus sieht. Ein typisches Wortarten-Tagset umfasst zwischen 50 und 150 verschiedene Tags[12].

Als Standard für deutschsprachige Korpora hat sich das *Stuttgart-Tübingen Tagset* (kurz: STTS) durchgesetzt[13]. Das sogenannte *kleine Tagset*, das keine expliziten Tags für Flexionsmorphologie enthählt, umfasst 54 Tags. Neben der Wortklasse werden weitere Eigenschaften wie die syntaktische Position bzw. Distribution des Wortes, seine grammatische Funktion und morphologische oder semantische Eigenschaften berücksichtigt. Zusätzlich deckt das Tagset auch Elemente ab, die man gemeinhin gar nicht als Wort klassifizieren würde, die aber als Token in authentischer, geschriebener Sprache vorkommen wie Satz- und andere Sonderzeichen. Die Verwendung der Tags wird im Annotationsschema bzw. den *Annotationsrichtlinien* (engl. ‚Guidelines', auch Tagging-Guidelines) beschrieben[14].

Wortartentags basieren auf einer Mischung unterschiedlicher Kategorisierungen. Im Folgenden illustrieren wir dies anhand von Tags des STTS.

---

[10] *Part of Speech* ist die englische Bezeichnung für Wortart.

[11] Vgl. Leech und Wilson (1996), S. 3.

[12] Siehe z.B. Schmid (2008).

[13] Ob es ein Zufall ist, dass die Nachnamen der vier maßgeblichen Autorinnen in Stuttgart und Tübingen – Anne Schiller, Christine Thielen, Simone Teufel und Christine Stöckert – ebenfalls zu ‚STTS' abgekürzt werden können?
Für historische Sprachstufen des Deutschen wurde das Tagset *HiTS* entworfen. Dipper et al. (2013) vergleichen HiTS mit STTS. Standards für das Englische sind die Varianten des CLAWS-Tagsets des British National Corpus (BNC) und das Penn Treebank Tagset, vgl. McEnery und Wilson (2001).

[14] Die Begriffe *Annotationsschema* und *Annotationsrichtlinien* werden austauschbar verwendet. Von der wörtlichen Bedeutung her steht bei Schema die Beschreibung der Kategorien im Vordergrund und bei den Richtlinien Handlungsanweisungen an die Annotatoren. In der Praxis wird diese Unterscheidung aber nicht durchgeführt.

- **Distribution** (d.h. positionelle Eigenschaften): z.B. Präposition versus Postposition

  (7) *APPR*: Die Zuschauer standen *entlang* der Straße.

  (8) *APPO*: Die Zuschauer standen die ganze Straße *entlang*.

- **Syntaktische Funktion**: z.B. attributiv versus prädikativ verwendetes Adjektiv

  (9) *ADJA*: Die *damaligen* Probleme sind uns heute nicht fremd.

  (10) *ADJD*: Damit waren sie *quitt*.

- **Morphologische Merkmale**: z.B. finite versus nicht-finite Verbform

  (11) *VVFIN*: Er *schreibt* Tagebuch.

  (12) *VVPP*: Er hat Tagebuch *geschrieben*.

  (13) *VVINF*: Er versuchte, Tagebuch zu *schreiben*.

- **Semantische Merkmale**: z.B. Appellativum ('Normales Nomen') versus Eigenname

  (14) *NN*: Verkleidete *Fischer* jagen nackte Amerikaner.

  (15) *NE*: Bundesaußenminister *Fischer* stimmte zu.

Oft unterscheiden sich die Wortarten in mehr als einem Merkmal. Finite und nicht-finite Verben zum Beispiel unterscheiden sich nicht nur morphologisch, sondern auch in ihrer syntaktischen Distribution: Nur finite Verben treten in der Verb-Zweit-Position auf; Partizipien werden zusammen mit Hilfsverben verwendet, reine Infinitive hingegen mit Modalverben usw. Ebenso unterscheiden sich attributiv und prädikativ verwendete Adjektive nicht nur in der Distribution, sondern auch in den morphologischen Merkmalen. Nur erstere werden flektiert und kongruieren mit dem begleitenden Nomen in Numerus, Genus und Kasus.

Das Stuttgart-Tübingen Tagset (STTS)

Dem STTS wurden als wichtigster Gliederungsaspekt distributionelle Kriterien zu Grunde gelegt. Bei Artikeln wird daher z.B. nicht nach definitem und indefinitem Artikel (*der*, *ein*) unterschieden, „da sie sich distributionell betrachtet gleich verhalten"[15]. Beide Artikelformen erhalten das Tag *ART*. Eine andere grundlegende Entscheidung bei der Entwicklung des STTS war, dass jede Wortform eines Textes genau einen Tag erhalten soll. Als Konsequenz davon werden Teile von Mehrwortlexemen unabhängig voneinander annotiert[16]. Bei der Vergabe der Tagnamen wurde auf das Prinzip der Teilbarkeit geachtet[17]:

---

[15] Vgl. Schiller et al. (1999), S. 33.

[16] Im englischen BNC werden in diesem Fall *ditto tags* vergeben, vgl. McEnery und Wilson (2001), S. 50. Jedes Token eines Mehrwortlexems erhält die Wortart des Gesamtausdrucks gefolgt von zwei Ziffern: der Gesamtzahl der Einzeltoken des komplexen Ausdrucks und dem jeweiligen Rang des gegebenen Tokens. *All of a sudden* (ganz plötzlich) wird zum Beispiel zu *all RR41 of_RR42 a RR43 sudden_RR44*, wobei *RR* das Label für *Adverb* ist.

[17] Vgl. McEnery und Wilson (2001), S. 51.

> Das Tagset ist hierarchisch strukturiert. (...) die *tags* bestehen aus möglichst selbsterklärenden Buchstabensequenzen, die von links nach rechts gelesen zuerst die Hauptwortart und dann die Unterwortart kodieren, also von der allgemeineren Information zur spezifischeren hinführen. (Schiller et al., 1999, S. 4)

Die Klasse der Pronomen *P* wird am stärksten unterteilt, was sich auch in den zusammengesetzten Tagnamen widerspiegelt. Je nach Funktion werden sie zu *D* (Demonstrativ), *I* (Indefinit), *PER* (PERsonal), *POS* (POSsessiv), *REL* (RELativ), *RF* (ReFlexiv), *W* (interrogativ oder relativ) oder *AV* (AdVerbial). Zusätzlich werden die meisten Pronomen noch nach ihrer Distribution spezifiziert: *S* (Substituierend) bzw. *AT* (ATtribuierend). Ganz systematisch entstehen so die Tagnamen, z.B. *PPOSS* steht für ein Pronomen, *POS*sessiv, Substituierend und *PPOSAT* für ein Pronomen, *POS*sessiv, *AT*tribuierend. Ein Ausschnitt des kleinen STTS-Tagsets wird in Tabelle 3 mit Beispielen illustriert[18].

Das Tagset als Balanceakt

Ein Tagset stellt immer einen Kompromiss dar zwischen Genauigkeit und Handhabbarkeit. Im STTS werden zum Beispiel prädikativ und adverbial verwendete Adjektive nicht unterschieden, sondern zur gemeinsamen Klasse ADJD zusammengefasst. Der Grund dafür ist, dass ein automatischer Tagger hier viele Fehler machen würde, weil zur Disambiguierung oft der gesamte Satz analysiert werden müsste. Die Verwendung eines unterspezifizierten Tags ist in diesem Fall gut zu vertreten, weil fast alle prädikativ verwendeten Adjektive auch adverbial auftreten können und umgekehrt. Für die wenigen Ausnahmen, die nur in einer der beiden Verwendungsweisen vorkommen können (ggf. zusätzlich zur attributiven Verwendung)[19], wie *untertan* (nur prädikativ) oder *ständig* (nicht prädikativ), geht diese Information allerdings verloren.

Ein ähnlicher Fall liegt bei den Verben *haben*, *sein* und *werden* vor, die neben ihren Hilfsverbfunktion zur Bildung von Perfekt, Futur oder Passiv auch als Vollverben auftreten können (*haben* im Sinne von *besitzen*, *sein* und *werden* als Kopula). Sie werden gemäß STTS immer als Auxiliar (Hilfsverb) gekennzeichnet, unabhängig davon, ob sie im konkreten Fall als Voll- oder Hilfsverb verwendet werden. Auch hier würden viele automatische Tagger bei der Disambiguierung scheitern, da wegen der Verbstellungsvarianten im Deutschen oft nur eine Analyse des gesamten Satzes ausreichend Informationen zur Auflösung liefern würde. Bei vielen Tagsets gibt es Kompromisse dieser Art, die in Hinblick auf die automatische Vorverarbeitbarkeit gemacht werden.

Manche Unterscheidungen sind auch für die Annotatoren schwierig. Ist *VW* in den beiden folgenden Beispielen ein Eigenname oder ein normales Nomen?

(16) Spontane Streiks bei *VW* in Emden.

(17) Wir hatten einen *VW* besessen.

Ist *gelehrt* in den beiden nächsten Beispielen ein Adjektiv oder ein verbales Partizip? Hier besteht eine Ambiguität zwischen der Kopulakonstruktion mit prädikativem *ADJD* und der Passivkonstruktion mit verbalem *VVPP*.

---

[18] Vgl. Schiller et al. (1999), S. 6f.

[19] Siehe auch Duden, Bd. 4 *Die Grammatik*, § 450ff.

- **Distribution** (d.h. positionelle Eigenschaften): z.B. Präposition versus Postposition

  (7) *APPR*: Die Zuschauer standen *entlang* der Straße.

  (8) *APPO*: Die Zuschauer standen die ganze Straße *entlang*.

- **Syntaktische Funktion**: z.B. attributiv versus prädikativ verwendetes Adjektiv

  (9) *ADJA*: Die *damaligen* Probleme sind uns heute nicht fremd.

  (10) *ADJD*: Damit waren sie *quitt*.

- **Morphologische Merkmale**: z.B. finite versus nicht-finite Verbform

  (11) *VVFIN*: Er *schreibt* Tagebuch.

  (12) *VVPP*: Er hat Tagebuch *geschrieben*.

  (13) *VVINF*: Er versuchte, Tagebuch zu *schreiben*.

- **Semantische Merkmale**: z.B. Appellativum ('Normales Nomen') versus Eigenname

  (14) *NN*: Verkleidete *Fischer* jagen nackte Amerikaner.

  (15) *NE*: Bundesaußenminister *Fischer* stimmte zu.

Oft unterscheiden sich die Wortarten in mehr als einem Merkmal. Finite und nicht-finite Verben zum Beispiel unterscheiden sich nicht nur morphologisch, sondern auch in ihrer syntaktischen Distribution: Nur finite Verben treten in der Verb-Zweit-Position auf; Partizipien werden zusammen mit Hilfsverben verwendet, reine Infinitive hingegen mit Modalverben usw. Ebenso unterscheiden sich attributiv und prädikativ verwendete Adjektive nicht nur in der Distribution, sondern auch in den morphologischen Merkmalen. Nur erstere werden flektiert und kongruieren mit dem begleitenden Nomen in Numerus, Genus und Kasus.

Das Stuttgart-Tübingen Tagset (STTS)

Dem STTS wurden als wichtigster Gliederungsaspekt distributionelle Kriterien zu Grunde gelegt. Bei Artikeln wird daher z.B. nicht nach definitem und indefinitem Artikel (*der, ein*) unterschieden, „da sie sich distributionell betrachtet gleich verhalten"[15]. Beide Artikelformen erhalten das Tag *ART*. Eine andere grundlegende Entscheidung bei der Entwicklung des STTS war, dass jede Wortform eines Textes genau einen Tag erhalten soll. Als Konsequenz davon werden Teile von Mehrwortlexemen unabhängig voneinander annotiert[16]. Bei der Vergabe der Tagnamen wurde auf das Prinzip der Teilbarkeit geachtet[17]:

---

[15] Vgl. Schiller et al. (1999), S. 33.

[16] Im englischen BNC werden in diesem Fall *ditto tags* vergeben, vgl. McEnery und Wilson (2001), S. 50. Jedes Token eines Mehrwortlexems erhält die Wortart des Gesamtausdrucks gefolgt von zwei Ziffern: der Gesamtzahl der Einzeltoken des komplexen Ausdrucks und dem jeweiligen Rang des gegebenen Tokens. *All of a sudden* (*ganz plötzlich*) wird zum Beispiel zu *all_RR41 of_RR42 a_RR43 sudden_RR44*, wobei *RR* das Label für *Adverb* ist.

[17] Vgl. McEnery und Wilson (2001), S. 51.

> Das Tagset ist hierarchisch strukturiert. (...) die *tags* bestehen aus möglichst selbsterklärenden Buchstabensequenzen, die von links nach rechts gelesen zuerst die Hauptwortart und dann die Unterwortart kodieren, also von der allgemeineren Information zur spezifischeren hinführen. (Schiller et al., 1999, S. 4)

Die Klasse der Pronomen *P* wird am stärksten unterteilt, was sich auch in den zusammengesetzten Tagnamen widerspiegelt. Je nach Funktion werden sie zu *D* (Demonstrativ), *I* (Indefinit), *PER* (PERsonal), *POS* (POSsessiv), *REL* (RELativ), *RF* (ReFlexiv), *W* (interrogativ oder relativ) oder *AV* (AdVerbial). Zusätzlich werden die meisten Pronomen noch nach ihrer Distribution spezifiziert: *S* (Substituierend) bzw. *AT* (ATtribuierend). Ganz systematisch entstehen so die Tagnamen, z.B. *PPOSS* steht für ein *P*ronomen, *POS*sessiv, *S*ubstituierend und *PPOSAT* für ein *P*ronomen, *POS*sessiv, *AT*tribuierend. Ein Ausschnitt des kleinen STTS-Tagsets wird in Tabelle 3 mit Beispielen illustriert[18].

Das Tagset als Balanceakt

Ein Tagset stellt immer einen Kompromiss dar zwischen Genauigkeit und Handhabbarkeit. Im STTS werden zum Beispiel prädikativ und adverbial verwendete Adjektive nicht unterschieden, sondern zur gemeinsamen Klasse ADJD zusammengefasst. Der Grund dafür ist, dass ein automatischer Tagger hier viele Fehler machen würde, weil zur Disambiguierung oft der gesamte Satz analysiert werden müsste. Die Verwendung eines unterspezifizierten Tags ist in diesem Fall gut zu vertreten, weil fast alle prädikativ verwendeten Adjektive auch adverbial auftreten können und umgekehrt. Für die wenigen Ausnahmen, die nur in einer der beiden Verwendungsweisen vorkommen können (ggf. zusätzlich zur attributiven Verwendung)[19], wie *untertan* (nur prädikativ) oder *ständig* (nicht prädikativ), geht diese Information allerdings verloren.

Ein ähnlicher Fall liegt bei den Verben *haben*, *sein* und *werden* vor, die neben ihren Hilfsverbfunktion zur Bildung von Perfekt, Futur oder Passiv auch als Vollverben auftreten können (*haben* im Sinne von *besitzen*, *sein* und *werden* als Kopula). Sie werden gemäß STTS immer als Auxiliar (Hilfsverb) gekennzeichnet, unabhängig davon, ob sie im konkreten Fall als Voll- oder Hilfsverb verwendet werden. Auch hier würden viele automatische Tagger bei der Disambiguierung scheitern, da wegen der Verbstellungsvarianten im Deutschen oft nur eine Analyse des gesamten Satzes ausreichend Informationen zur Auflösung liefern würde. Bei vielen Tagsets gibt es Kompromisse dieser Art, die in Hinblick auf die automatische Vorverarbeitbarkeit gemacht werden.

Manche Unterscheidungen sind auch für die Annotatoren schwierig. Ist *VW* in den beiden folgenden Beispielen ein Eigenname oder ein normales Nomen?

(16) Spontane Streiks bei *VW* in Emden.

(17) Wir hatten einen *VW* besessen.

Ist *gelehrt* in den beiden nächsten Beispielen ein Adjektiv oder ein verbales Partizip? Hier besteht eine Ambiguität zwischen der Kopulakonstruktion mit prädikativem *ADJD* und der Passivkonstruktion mit verbalem *VVPP*.

---

[18] Vgl. Schiller et al. (1999), S. 6f.

[19] Siehe auch Duden, Bd. 4 *Die Grammatik*, § 450ff.

(18) Er ist gelehrt.

(19) Hier wird Linguistik gelehrt.

Wortart	Beschreibung	Beispiele (unterstrichen)
ADJA	attributives Adjektiv	*das große Haus*
ADJD	adverbiales oder prädikatives Adjektiv	*er fährt schnell*; *er ist schnell*
ADV	Adverb	*schon*; *bald*; *doch*
APPR	Präposition oder Zirkumposition links	*in der Stadt*; *ohne mich*; *um ihn herum*
APPO	Postposition	*ihm zufolge*; *der Sache wegen*
KON	nebenordnende Konjunktion	*und*; *oder*; *aber*
KOKOM	Vergleichskonjunktion	*als*; *wie*
NN	normales Nomen	*Tisch*; *Herr*; *das Reisen*
NE	Eigennamen	*Hans*; *Hamburg*; *HSV*
PDS	substituierendes Demonstrativpronomen	*dieser*; *jener*
PDAT	attribuierendes Demonstrativpronomen	*jener Mensch*
PIS	substituierendes Indefinitpronomen	*keiner*; *viele*; *man*; *niemand*
PIAT	attribuierendes Indefinitpronomen	*kein Mensch*; *irgendein Glas*
PPER	irreflexives Personalpronomen	*ich*; *er*; *ihm*; *mich*; *dir*
PPOSS	substituierendes Possessivpronomen	*meins*; *deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein Buch*; *deine Mutter*
PRELS	substituierendes Relativpronomen	*der Hund, der*
PRELAT	attribuierendes Relativpronomen	*der Mann, dessen Hund*
PRF	reflexives Pronomen	*sich*; *einander*; *dich*; *mir*
PWS	substituierendes Interrogativpronomen	*wer*; *was*
PWAT	attribuierendes Interrogativpronomen	*welche Farbe*; *wessen Hut*
PWAV	adverbiales Interrogativ- oder Relativpronomen	*warum*; *wo*; *wann*; *worüber*; *wobei*
VVFIN	finites Vollverb	*du gehst*; *wir kommen an*
VVINF	Infinitiv eines Vollverbs	*er will gehen*; *ankommen*
VVPP	Partizip Perfekt eines Vollverbs	*hat getroffen*; *sie sind entlaufen*
VAFIN	finites Hilfsverb (Auxiliar)	*du bist*; *wir werden*
VMFIN	finites Modalverb	*wir müssen gehen*
$.	satzbeendende Interpunktion	. ? ! ; :

Tabelle 3: Ausschnitt aus dem Stuttgart-Tübingen Tagset (STTS)

Mit systematischen Ambiguitäten wie in den Beispielen (16) – (19) wird sehr unterschiedlich umgegangen. Im *British National Corpus*, dem britischen Referenzkorpus, sind sogenannte *Portmanteau-Tags* erlaubt, die aus einer Kombination von zwei Tags bestehen, zum Beispiel `heard&VVD-VVN`; zeigt, dass das Token *heard* entweder in der

einfachen Vergangenheit (*VVD*) oder als Partizip Perfekt (*VVN*) verwendet wird. Die Zeichen & und ; markieren die Grenzen der Teiltags (im TEI-Format[20]).

Um zwischen Eigennamen und ‚normalen Nomen' unterscheiden zu können, definieren die STTS-Guidelines eine in sich abgeschlossene Liste von Eigennamen-Unterklassen wie *Vorname*, *Nachname* und *Firmenname* und nur diese werden als Eigenname ‚NE' getaggt[21]. In anderen Fällen, wie bei der ADJD-VVPP-Ambiguität (vgl. Beispiel (18) und (19)) geben die Richtlinien linguistische Entscheidungshilfen und listen zusätzlich bereits bekannte, lexikalisierte ADJD-Formen auf.

Linguistische Kriterien: ADJD vs. VVPP[22]:

1. Kann der Satz ins Aktiv gesetzt werden mit gleicher Semantik?
   Ja → VVPP
2. Gibt es eine *von*-PP oder ähnliche PP, die auf Verbsemantik hinweist?
   Ja → VVPP
3. Ist eine Ersetzung durch ein semantisch nahes Adjektiv möglich?
   Ja → ADJD

Die linguistischen Kriterien stellen einen geordneten Fragenkatalog dar. Man beginnt mit Frage 1, nur wenn diese mit „nein" beantwortet wird, geht man zu Frage 2 weiter. Formal gesehen handelt es sich hier um einen *Entscheidungsbaum*. Im Folgenden wenden wir diese Kriterien auf die Beispiele (18) und (19) auf der vorhergehenden Seite an.

(20) Er ist gelehrt.
   1.*Sie lehrt ihn.
   2.*Er ist von ihr gelehrt.
   3. Er ist klug.
   → ADJD

(21) Hier wird Linguistik gelehrt.
   1. Sie lehrt hier Linguistik.
   2. Hier wird Linguistik von ihr gelehrt.
   3.*Hier wird Linguistik klug.
   → VVPP

Morphologie und Lemmatisierung

Die Annotation von Flexionsmorphologie wird oft vom reinen Wortarten-Tagging unterschieden. Hierzu wird das Token analysiert und auf seine Grundform, das Lemma,

---

[20] Vgl. Leech und Wilson (1996), S. 17, McEnery und Wilson (2001). Im BNC werden die Portmanteau-Tags nur für Ambiguitäten verwendet, die für einen automatischen Tagger schwer aufzulösen sind.

[21] Der Ausdruck *VW* als Firmenname in Beispiel (16) wird gemäß STTS als NE getaggt. In seiner Verwendung als Produktbezeichnung in Beispiel (17) wird er hingegen als normales Nomen ‚NN' bezeichnet, da Produktnamen in den STTS-Guidelines nicht in der Liste der Eigennamen aufgeführt werden.

[22] Vgl. Schiller et al. (1999), S. 24.

zurückgeführt. Dabei erhält man eine morphologische Analyse, die auf ein morphologisches Tagset abgebildet werden kann.

Flexionsmorphologie umfasst Kategorien wie *Kasus*, *Genus*, *Numerus*, *Person*, *Tempus* und *Modus*. Das sogenannte *große Tagset des STTS* verwendet zusätzlich zu den genannten Kategorien auch noch die Kategorien *Grad* (steigerbar)[23], *Definitheit* und *Flexion*. Letzteres ist, wie in (22) und (23) dargestellt, die Markierung für stark (St), schwach (Sw) oder gemischt (Mix) flektierte Adjektive und Nomen (u.a. Nominalisierungen von Adjektiven)[24].

(22) a. *mit ganzem*/ADJA:Pos.Masc.Dat.Sg.**St** *Einsatz*
b. *mit dem ganzen*/ADJA:Pos.Masc.Dat.Sg.**Sw** *Hausrat*
c. *mit einem ganzen*/ADJA:Pos.Masc.Dat.Sg.**Mix** *Apfel*

(23) a. *ich Armer*/NN<.ADJ:Masc.Nom.Sg.**St** (deadjektivisch)
b. *der Beamte*/NN:Masc.Nom.Sg.**Sw**
c. *eine Rote*/NN|ADJ:Fem.Nom.Sg.**Mix** (deadjektivisch)
d. *die Kosten*/NN:*.Nom.Pl.

Kann ein morphologischer Wert nicht eindeutig zugewiesen werden, wird ein Sternchen vergeben, wie z.B. für das Genus bei *Kosten* in Beispiel (23). Manchmal müssen Kategorien aus technischen Gründen angegeben werden, obwohl sie nur bei einer Teilklasse vorhanden sind. Diese Kategorie wird dann durch einen Unterstrich symbolisiert. Das Nomen *Kosten* kann hier wieder als Beispiel dienen. Wie die Mehrzahl der Nomen wird es keiner Flexionsklasse zugeteilt und erhält daher an der entsprechenden Position in der Morphologie einen Unterstrich[25].

Durch die Kombination von Wortart und morphologischer Information wächst das sogenannte große STTS-Tagset auf mehrere hundert Elemente.

Exkurs: Tagging

Part-of-Speech Tagging bezeichnet die automatische Zuweisung von Wortartentags (*Part-of-Speech Tags*) zu einzelnen Wortformen. Es ist ein wichtiger Schritt in der Textaufbereitung und Grundlage für viele weiterführende Annotationen[26]. Automatische Methoden sind schon weit entwickelt und erreichen hohe Genauigkeiten (95% bis 98% Pro-Wort-Akkuratheit)[27]. Das folgende Schaubild gibt eine (vereinfachte) schematische Übersicht über die wichtigsten Komponenten des Taggings.

---

[23] *Grad* hat die Werte Positiv (Pos), Komparativ (Comp) und Superlativ (Sup).
[24] Vgl. Schiller et al. (1999), S. 13, 20.
[25] Vgl. Schiller et al. (1999), S. 8.
[26] In der Computerlinguistik dient Text, der mit Wortartentags annotiert ist, als Datengrundlage für viele Anwendungen, z.B. bei der Informationsextraktion, Sprachsynthese, Computerlexikographie oder Termextraktion.
[27] Vgl. Schmid (2008).

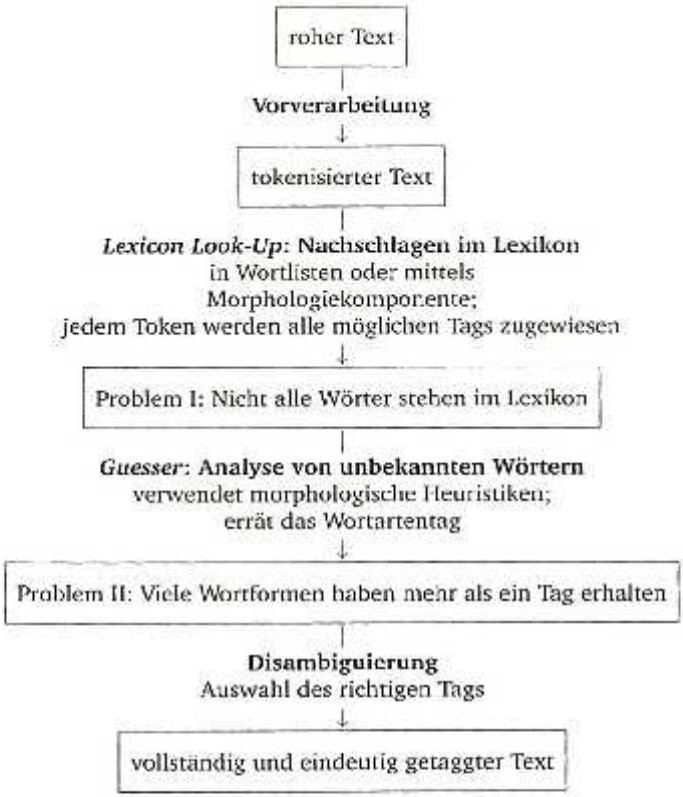


Abbildung 1: Schematische Darstellung des Part-of-Speech Taggings

Unter *Lexikon* versteht man hier eine Auflistung der Wortformen mit jeweils allen möglichen Lesarten, wie z.B. die Einträge von *einen* und *einende* in einem Lexikon, das vom TreeTagger[28] verwendet wird. Zur besseren Lesbarkeit ist hier bei jedem Lexikoneintrag die Wortform unterstrichen:

```
einen     ART.Akk ein     INDEF.subst.Akk ein  VVFIN einen  VVINF einen
einende   ADJ.Akk einend  ADJ.Nom einend
```

Eine „Lesart" wird hier jeweils durch ein Paar bestehend aus Wortartentag und entsprechendem Lemma repräsentiert, z.B. *ART.Akk ein* oder *INDEF.subst.Akk ein*.

[28] Vgl. Schmid (1995).

Die Disambiguierung ist neben dem Raten von Tags für unbekannte Wortformen das größte Problem beim Taggen. Automatische Tagger können danach klassifiziert werden, wie sie dieses Problem lösen.

- *Symbolische Tagger* verwenden (meist) handgeschriebene Regeln wie ‚Wenn ein Wort zwischen Artikel- und Verbiesart ambig ist (z.B. *einen*), dann wähle das Verb-Tag, wenn das vorangehende Wort zu ist'. Der Tagger TAGGIT [29], der zum Taggen des BROWN Corpus eingesetzt wurde, basiert zum Beispiel auf Kontextmuster-Regeln, weist 71 verschiedene Tags zu und verwendet zur Disambiguierung 3 300 Disambiguierungsregeln[30].
- *Stochastische Tagger* werden *trainiert*, indem sie die Frequenzen von Wörtern und Tags eines vorannotierten Trainingskorpus zählen[31] und daraus Wahrscheinlichkeiten ableiten:
  - lexikalische Wahrscheinlichkeit: das wahrscheinlichste Tag für ein Token (z.B. *einen* ist eher ein Artikel als ein Verb)
  - kontextuelle Wahrscheinlichkeit: das wahrscheinlichste Tag für ein Token in einen bestimmten Kontext (d.h. einer Sequenz von vorangehenden oder nachfolgenden Tags und Wörtern, z.B. *einen* vor einem Satzendepunkt ist eher ein Verb als ein Pronomen)

  Die entscheidende Aufgabe des Taggers besteht darin, die optimale Balance zwischen diesen beiden Ergebnissen zu finden. Beispiele für stochastische Tagger sind der TnT-Tagger[32] oder der TreeTagger[33].
- *Hybride Tagger* verbinden symbolische Regeln mit stochastischen, korpusbasierten Methoden. Sie ‚lernen' die Gewichtung der Regeln anhand ihrer Anwendung auf Korpusdaten und anschließendem Vergleich der Ergebnisse mit einem vorannotierten Korpus (einem *Goldstandard*). Ein prominenter Vertreter dieser Methode ist der Brill-Tagger[34], der neben den Wahrscheinlichkeiten auch symbolische Regeln lernt. Zunächst wird aus dem Goldstandard für jedes Token das wahrscheinlichste Tag abgeleitet. Im ersten Taggingschritt wird jedem Token in dem zu annotierenden Text einfach nur sein wahrscheinlichstes Tag zugeordnet. Das so getaggte Korpus wird mit der Annotation des Goldstandards verglichen. Natürlich gibt es viele Abweichungen, immer dann, wenn ein Token im Goldstandard nicht mit seinem wahrscheinlichsten Tag auftritt, sondern mit einem anderen, weniger wahrscheinlichen. Dieser erste Abgleich ist der Ausgangspunkt (oder auch die *Baseline*) für das weitere Training. Der Tagger muss versuchen, ein besseres Ergebnis zu erzielen. Er ruft eine Liste von Reparaturregeln (*Transformationsregeln*) auf, die versuchsweise einzelne Tags kontextabhängig ersetzen. Das geänderte Korpus wird wieder mit dem Goldstandard

[29] Vgl. Greene und Rubin (1971).
[30] Vgl. McEnery und Wilson (2001).
[31] Es gibt auch Methoden, Tag-Wahrscheinlichkeiten auf nicht-annotierten Trainingskorpora zu schätzen; siehe dazu allgemein Manning und Schütze (1999, Kap. 10), Jurafsky und Martin (2008, Kap. 5,6).
[32] Vgl. Brants (2000).
[33] Vgl. Schmid (1995).
[34] Vgl. Brill (1995).

verglichen. Ist das Resultat besser als die Baseline, werden die Regeln übernommen, ansonsten werden sie verworfen. Drei auf diese Art gelernte Regeln für das Deutsche sind z.B. die folgenden (die Tags stammen aus dem STTS-Tagset. Die zweite Zeile ist jeweils eine umgangssprachliche Umschreibung der Regel)[35].

(24) `ART PRELS PREVTAG $,`
= Ersetze ART durch PRELS, wenn vorher das Tag $, steht.

(25) `PTKZU APPR NEXT1OR2OR3TAG NN`[36]
= Ersetze PTKZU durch APPR, wenn innerhalb der nachsten 3 Tags NN kommt.

(26) `ART PDS WDNEXTTAG das ADV`
= Ersetze ART durch PDS, wenn das aktuelle Wort *das* heißt und der Tag danach ADV ist.

Der Brill-Tagger versucht, auch auf der Wortbildungsebene Regeln zu lernen. Eine automatisch aus dem Korpus abgeleitete Regel ist z.B. die tatsächlich auch linguistisch motivierte Aussage (hier in verständlicher Umschreibung wiedergegeben):

(27) Bei Präfix *um-* ersetze VVPP durch ADJD.

4.3.2 Syntaktische Annotation

Die nächste Ebene der Annotation ist die Syntax im Sinne wortübergreifender Analyse. Korpora mit syntaktischer Annotation nennt man auch *Baumbanken*[37]. Die Bezeichnung hat ihren Ursprung darin, dass die ersten syntaktischen Annotationsvorhaben strukturelle Bäume als Analyseform vorsahen.

Graphenstruktur

Ein Baum hat normalerweise einen eindeutigen Wurzelknoten an der Spitze (‚root node')[38], der über der gesamten Wortkette steht. In Beispiel (29) auf S. 72 ist das der VP-Knoten. Ein Baum verzweigt sich wohlgeordnet, so dass sich keine Äste (formaler ausgedrückt: *Kanten* ‚edges') überkreuzen und jeder Knoten (‚node') nur einen eindeutigen Mutterknoten besitzt – und nicht zwei oder mehrere. Möchte man Überkreuzungen zulassen (also *überkreuzende Kanten*), arbeitet man, wenn man es mathematisch genau

---

[35] Vielen Dank an Stefanie Dipper, die uns die die Beispielregeln zur Verfügung stellte. Der Brill-Tagger wurde hierzu auf 779 STTS-annotierten Sätzen des TIGER-Korpus plus 820 nichtannotierten Sätzen trainiert. Es reichten für das Deutsche insgesamt 100–200 Regeln aus, um mit einer Genauigkeit von 97% zu taggen.

[36] PTKZU = *zu* vor Infinitiv.

[37] Von Englisch ‚treebank'. Der Begriff wurde von Geoffrey Leech geprägt im Zusammenhang mit einem Vorgängerprojekt des englischen SUSANNE Korpus, vgl. Sampson (2003), S. 40, Fn. 1.

[38] Syntaxbäume wachsen verkehrt herum, mit der Wurzel nach oben.

nimmt, nicht mit Baumgraphen, sondern mit allgemeineren Graphenstrukturen[39]. Die Blätter des Baumes sind die *terminalen* Knoten (von der englischen Bezeichnung „terminal" für *abschließend, endständig*). Sie bezeichnen hier die einzelnen Wörter des Satzes. Alle Knoten außer den terminalen werden als *nicht terminale* Knoten bezeichnet, wobei die Knoten, die unmittelbar über den Wörtern stehen, auch *Präterminale* genannt werden. Im Beispiel sind dies die Knoten mit den Wortartentags. Als zusätzliche Ebene findet man in vielen Baumannotationen auch *sekundäre Kanten*, die nicht zur eigentlichen Baumstruktur gehören[40].

Dependenz und Konstituenz

Bei der syntaktischen Annotation unterscheidet man zwei grundlegende Modelle: die *Konstituentenstruktur* und die *Dependenzstruktur*. Zur Illustration der Unterschiede wollen wir Ihnen ein einfaches Beispiel geben (siehe auch die Darstellungen in (28) – (30)). Die Verbalgruppe *ein einfaches Beispiel geben* aus dem letzten Satz besteht aus vier Wörtern der Wortarten (gemäß STTS): ART ADJA NN VVINF. Die Wörter sind nicht ganz gleichberechtigt. Obwohl nur eines der vier Wörter ein Verb ist, bezeichnen wir die ganze Sequenz als *Verbalgruppe*. Wir heben das Verb *geben* als Kern (auch *Kopf*) der Sequenz hervor.

Sowohl der dependenzbasierte als auch der konstituentenbasierte Ansatz gehen von einer hierarchischen Strukturierung von Sätzen aus. Sie unterscheiden sich jedoch in Bezug auf die Elemente, die in der hierarchischen Gliederung geordnet werden: In einer Konstituentenstruktur sind es Konstituenten, also abstrakte Einheiten, die jeweils ein oder mehrere Wörter repräsentieren, z.B. Verbalphrase VP, Nominalphrase NP in Abb. (29). In der Dependenzstruktur beschränkt man sich auf die Wörter selbst, vgl. Abb. (30).

Die Konstituentenstrukturanalyse geht auf den amerikanischen Strukturalismus zurück[41]. Man nimmt an, dass Sätze aus hierarchisch geschachtelten Untereinheiten bestehen, die man zum Beispiel durch Klammerung markieren kann. Diese Untereinheiten sind Sequenzen von zusammenhängenden Wörtern, die als *Konstituenten* bezeichnet werden[42], vgl. (28) und (29). Beachten Sie, dass jede Wortform für sich genommen ebenfalls als Konstituente betrachtet werden kann. Ein prototypisches Beispiel für ein Korpus mit reiner Konstituentenanalyse ist die amerikanische *Penn Treebank*[43].

(28) Klammerstruktur:
[$_{VP}$ [$_{NP}$ [$_{ART}$ ein] [$_{ADJA}$ einfaches] [$_{NN}$ Beispiel]] [$_{VVINF}$ geben]]

---

[39] Im Zusammenhang dieses Buches wollen wir nicht weiter auf die Unterschiede eingehen und werden vereinfachend auch dann von *Bäumen* reden, wenn es im mathematischen Sinne keine sind. In der TIGER-Baumbank zum Beispiel kommen überkreuzende Kanten zum Einsatz.

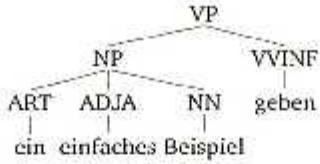
[40] In der TIGER-Baumbank werden sekundäre Kanten zum Beispiel verwendet, um geteilte Argumente in Koordinationen anzuzeigen, vgl. Abb. 7 auf S. 80.

[41] Ein wichtiger Vertreter ist Zellig Harris (1951). Es gab aber schon Vorläufer, vgl. Langer (2010).

[42] Konstituenten können durch Tests identifiziert werden (z.B. durch Ersetzung, Verschiebung oder Koordination), vgl. z.B. Pittner und Berman (2013) oder Klenk (2003).

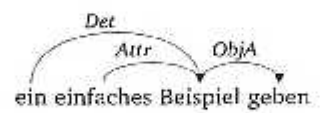
[43] Vgl. Marcus et al. (1993).

(29) Baumstruktur:



Ein wichtiger Vertreter der Dependenztheorie ist Lucien Tesnière[44]. In einer Dependenzanalyse besteht die Satzhierarchie aus Abhängigkeiten (*Dependenzen*) von Wörtern untereinander. Die Dependenzen werden durch Verknüpfungen von jeweils zwei Wörtern modelliert. Grafisch sind es Kanten eines Baums (bei Tesnière ‚connexions'). Die Verknüpfungen sind immer gerichtet. Genauer gesagt, gibt es immer ein *Regens* und ein davon abhängiges *Dependens*, vgl. Beispiel (30). *Geben* regiert *Beispiel*, welches wiederum *ein* und *einfaches* regiert[45]. Normalerweise stehen die abhängigen Elemente in einer bestimmten *grammatischen Funktion* zum Regens, im Beispiel sind es *Det(erminator)*, *Attr(ibut)* und *Akkusativobjekt* (*ObjA*)[46]. Obwohl eine Dependenzanalyse nicht zwingend die Angabe grammatischer Funktionen einschließt, sind beide Konzepte doch sehr eng miteinander verbunden. Man spricht auch von einer *funktionalen Analyse*. Ein prototypisches Beispiel für ein Korpus mit Dependenzannotationen ist die tschechische *Prague Dependency Treebank*[47].

(30) Funktionale Dependenzstruktur:



Hybride Modelle

Eine Konstituentenstruktur bildet zunächst nur syntaktische Kategorien ab und keine Funktionen. In vielen Projekten wird daher eine gemischte Repräsentation bevorzugt (*hybrides Modell*). Als Grundgerüst werden strukturelle Kategorien gebildet, die mit funktionalen Informationen angereichert werden. In einer Baumdarstellung kann man z.B. die Kategorien als Knotenlabel repräsentieren und die verbindenden Kanten mit funktionalen Labeln versehen. Wir verwenden hier dieselben Label wie im Dependenzbeispiel oben. Die Kerne (Köpfe) der VP und NP sind zusätzlich als *H(ea)d* markiert.

---

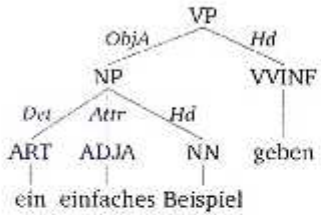
[44] Vgl. Tesnière (1959). Für eine Einführung siehe z.B. Weber (1997).

[45] In der grafischen Darstellung weisen die Pfeilspitzen normalerweise auf das Regens. Manche Korpora weichen allerdings von dieser Konvention ab.

[46] Die Funktionen können wie hier als Kantentags dargestellt werden.

[47] Auf der sogenannten *analytischen Ebene* der Annotation sind in der Prague Dependency Treebank reine Dependenzstrukturen annotiert, vgl. `ufal.mff.cuni.cz/pdt2.0/`.

(31) Hybride Baumstruktur:



Viele der Baumbanken, die eine konstituentenbasierte Grundarchitektur besitzen, fallen in die Klasse der hybriden Modelle, weil sie auf die auch funktionale Information darstellen. In der weiter vorne erwähnten Penn Treebank, die zunächst auf einem rein konstituentenbasierten Modell aufsetzte, wurde schlussendlich ein hybrides Annotationsschema umgesetzt, das z.B. vorsieht, dass Subjekte und adverbiale Präpositionalphrasen mit zusätzlichen funktionalen oder semantischen Tags auszeichnet werden, s. auch Abschnitt 4.3.3[48].

Phrasen und Chunks

Wir haben bisher die Begriffe Konstituente und Phrase fast wie Synonyme behandelt. Wenn man es genau nehmen möchte, dann ist eine Konstituente die kategorieneutrale Beschreibung einer *Phrase*. Letztere ist immer einer bestimmten Kategorie zugeordnet, z.B. Verbalphrase oder Nominalphrase. Man unterscheidet dabei sogenannte endozentrische und exozentrische Phrasen (siehe die Beispiele in (32)). Bei *endozentrischen Phrasen* existiert ein phraseninterner *Kopf*, welcher die kategoriellen Eigenschaften bestimmt, z.B. das Nomen in der Nominalphrase. Die sogenannten Projektionen des Kopfes sind bis zur maximalen, also der phrasalen Ebene von derselben Kategorie, hier im Beispiel sind sie nominal. Sie unterscheiden sich lediglich in der Projektionsebene (ausgedrückt durch Striche[49] oder Nummerierung, z.B. *N'* oder *N1*). Die maximale Ebene wird dann mit einem phrasalen Tag gekennzeichnet, hier *NP*. Bei einer *exozentrischen Phrase* ist der Mutterknoten von einem anderen kategoriellen Typ als alle seine Töchter. Hier werden verschiedene Phrasen zu einer funktionalen Einheit zusammengefasst, z.B. der Satzknoten *S*, der in traditionellen Analysen über der Subjekts-NP und der VP steht. Eine Formalisierung erfährt der Phrasenbegriff zum Beispiel durch die X-Bar-Struktur[50].

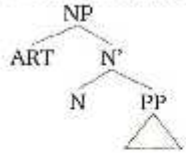
---

[48] Vgl. Marcus et al. (1993) und Marcus et al. (1994).

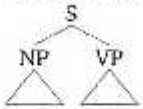
[49] Ursprünglich wurden die Striche als Oberstriche (englisch bars) gesetzt: $\bar{N}$.

[50] Siehe Jackendoff (1977). Das X-Bar-Schema findet in der Korpusannotation kaum Anwendung. Das hat zwei Gründe. Zum einen versucht man, Annotationen meistens möglichst theorieneutral zu halten – es sei denn, man plant explizit eine theoriebasierte Baumbank zu erstellen, wie z.B. die HPSG-basierte bulgarische BulTreebank (Simov und Osenova, 2003). Zum zweiten erzeugen X-Bar-Strukturen sehr schnell sehr große Bäume, was für den Annotationsvorgang und beim späteren Browsen durch das annotierte Korpus hinderlich ist.

(32) a. Endozentrische Phrase NP



b. Exozentrische Phrase S



Ein alternatives Konzept der syntaktischen Gruppierung sind *Chunks*. Das Konzept geht auf Steven Abney[51] zurück. Motiviert durch psycholinguistische Beobachtungen[52], definiert er „Brocken" (die wörtliche Übersetzung von ‚chunks'). Sie entsprechen prosodischen Einheiten, d.h. Sprecheinheiten, nach denen Sprecher intuitiv eine kleine Sprechpause einlegen. Wenn Laien einen Satz in Sprecheinheiten unterteilen sollen, tendieren sie dazu, Einheiten zu bilden, die genau solchen Chunks entsprechen.

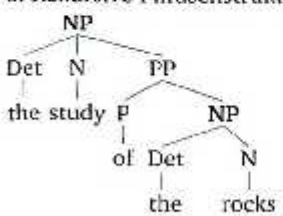
Definition 1 (Chunk, strikte Version). *Der nicht-rekursive Kernbereich einer Konstituente innerhalb eines Satzes, beginnend am Anfang der Konstituente bis hin zu ihrem (lexikalen) Kopf (nach Abney 1991).*

Das folgende Beispiel zeigt die Chunks eines englischen Satzes. Bei der Präpositionalgruppe ‚on his suitcase' trifft die Chunkdefinition dann zu, wenn man das Nomen ‚suitcase' anstelle der Präposition ‚on' als lexikalen Kopf der Gesamtstruktur betrachtet.

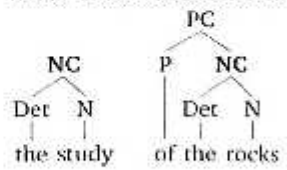
(33) [The bold man] [was sitting] [on his suitcase].

Eine besondere Eigenschaft von menschlicher Sprache ist, dass sie *rekursive* Strukturen aufweist, also in sich geschachtelte Einbettungen derselben Kategorie. In (34) weist z.B. die Baumstruktur links eine solche Einbettung auf: Sie enthält eine komplexe Nominalphrase (NP), bei der unter der maximalen NP eine weitere NP eingebettet ist. Beim nicht-rekursiven Chunking dagegen (vgl. die Struktur rechts) erhält man flache Analysen und keine rekursiven Einbettungen: Ergänzungen und Modifikatoren, die nach dem Kopf einer Phrase folgen, werden nicht in den Chunk des Kopfes eingeschlossen, sondern bilden einen unabhängigen Chunk. In der Struktur rechts bezeichnen die Tags *NC* und *PC* einen nominalen bzw. präpositionalen Chunk. Die Teilbäume von *the study* und *of the rocks* stehen als unabhängige Chunks nebeneinander.

(34) a. Rekursive Phrasenstruktur



b. Nicht-rekursives Chunking



[51] Vgl. Abney (1991).
[52] Vgl. Gee und Grosjean (1983).

Partielle und vollständige Analyse

Das Chunking (oder auch *Partial Parsing*) ist in der automatischen Sprachverarbeitung sehr verbreitet. Es erlaubt, Teilstrukturen mit relativ hoher Qualität zu analysieren, ohne dass man über die Gesamtstruktur des Satzes spekulieren muss. Dasselbe gilt für die Annotation von Korpora. Auch hier wird das Chunking eingesetzt als eigenständige Annotationsform oder auch als automatischer Vorverarbeitungsschritt einer vollständigen syntaktischen Analyse. Für das Deutsche wird die strenge Chunkdefinition nach Abney auf rekursive Strukturen erweitert, um Beispielen wie (35) gerecht zu werden, bei denen im pränominalen Bereich – anders als im Englischen – erweiterte Adjektivphrasen auftreten, hier z.B. die Adjektivgruppe *durch Fehlentscheidungen hochverschuldete* bei der das Adjektiv *hochverschuldete* durch die Präpositionalphrase *durch Fehlentscheidungen* erweitert ist[53].

(35) [NC die [AC [PC durch [NC Fehlentscheidungen]] [AC hochverschuldete]] Bahn]

Ein Beispiel für ein gechunktes (d.h. syntaktisch „partiell analysiertes") Korpus ist das *Tübinger Partiell Geparste Korpus des Deutschen / Schriftsprache* (kurz: TüPP-D/Z).

Repräsentation der syntaktischen Annotation

Wie sieht die syntaktische Annotation nun in der Praxis aus? Um einen Eindruck davon zu vermitteln, stellen wir drei syntaktische Tagsets beispielhaft an einem Satz vor.

Dependenzannotation. Als erstes betrachten wir ein Korpus, das an der Universität Hamburg im Rahmen eines Projekts zum automatischen Dependenzparsen erstellt wurde: Die *Hamburg Dependency Treebank*[54] umfasst mehr als vier Millionen manuell annotierte bzw. korrigierte Token, mehr als 100 000 Sätze. Zusätzlich enthält sie weitere, nur automatisch annotierte Sätze.

Tabelle 4 zeigt einen Teil des Tagsets für funktionale Dependenzen, das insgesamt aus 35 Tags besteht[55].

In Abb. 2 sehen Sie eine sehr einfache, grafische Darstellung der Dependenzstruktur des Satzes *Wir sind begeistert!*. Die Knoten des Baums entsprechen jeweils einem Token auf der Satzebene. Der oberste Knoten, der Wurzelknoten, ist hier ein Hilfskonstrukt ohne Entsprechung auf der Satzebene. Die Kanten haben sind hier ohne Pfeilspitzen dargestellt. Die Richtung der Abhängigkeit kann aus der relativen Knotenhöhe erschlossen werden: Der Knoten eines Regens ist höher dargestellt als der Knoten seines Dependens.

In unserem Beispiel markiert eine *S*(atz)-Kante vom abstrakten Wurzelknoten ausgehend das Wort *sind* als das eigentliche Wurzelwort des Satzes. Zwei weitere Kanten weisen auf *sind*. Sie verknüpfen über eine *SUBJ*(ekt)- bzw. eine *PRED*(ikativ)-Funktion, die abhängigen Knoten *wir* und *begeistert* mit ihrem Regens. Das Ausrufezeichen ist durch die „leere" Kante als unregiertes Element markiert.

---

[53] Das Beispiel stammt vereinfacht aus Müller (2004), S. 4, siehe ebenfalls Kermes (2003).

[54] Vgl. Foth et al. (2014), Korpus-Download: `https://corpora.uni-hamburg.de/drupal/de/islandora/object/treebank:hdt`.

[55] Vgl. Foth (2006).

Tag	Dependens	Regens
S	Wurzelwort eines Satzes (oder eines Satzfragments), normalerweise das finite Verb	Abstrakter Wurzelknoten 0
SUBJ	Kopfnomen eines Subjekts	finites Verb
PRED	nicht-verbales Prädikativ	Kopulaverb
AUX	Verb	Auxiliar
OBJA	Kopfnomen eines Akkusativobjekts	Verb
OBJD	Kopfnomen eines Dativobjekts	Verb
KOM	Vergleichswort (*als, wie*)	Bezugswort

Tabelle 4: Funktionale Tags der Hamburg Dependency Treebank

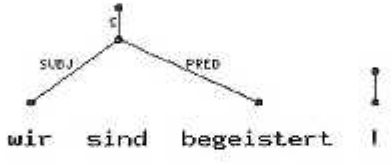


Abbildung 2: Dependenzannotation in der Hamburg Dependency Treebank

Der grafischen Baumdarstellung liegt eine Textdatei zugrunde, vgl. Abb. 3. Wenn Sie an den Details der textuellen Annotationsrepräsentation nicht interessiert sind, überspringen Sie den Rest dieses Paragraphens. Jeder kommagetrennte Block entspricht einer Dependenz und beginnt jeweils mit zwei Zahlen, die die Position des abhängigen Wortes im Satz angeben. Das erste Token *wir* z.B. nimmt die Position zwischen den Markierungen 0 und 1 ein, das zweite zwischen 1 und 2 usw. Die Wortart *cat* ist nach dem STTS getaggt, das Sie schon kennengelernt haben. Die syntaktische Information *SYN* gibt das funktionale Tag der Kante an sowie die Endposition des Regens, welches zusätzlich als Kommentar am Ende der Zeile ausbuchstabiert wird.

Konstituentenstruktur. Als Beispiele für phrasenstrukturelle Tagsets stellen wir die Annotationsschemata der TIGER-Baumbank und der beiden TüBa-Baumbanken (TüBa-D/S und TüBa-D/Z) vor.

Beide Tagsets umfassen je 25 nicht-terminale Tags. Der kleine Vergleich in Tabelle 5 weist schon auf gewisse Unterschiede hin: In TüBa werden topologische Felder annotiert[56], in der TIGER-Baumbank erhalten koordinierte Phrasen besondere Tags. In den TüBa-Baumbanken ist die Annotation von der Chunkidee beeinflusst, deshalb heißen z.B. nominale Konstituenten nicht *NP* sondern *NX*[57]. In der TIGER-Baumbank werden

[56] Vgl. Höhle (1986) bzw. Pittner und Berman (2013) für eine Einführung.

[57] Einen ausführlicheren Vergleich der beiden Annotationsschemata finden Sie in Ule und Hinrichs (2004) sowie bei Telljohann et al. (2004).

```
0 1 'Wir'
'cat' / 'PPER'
'SYN' -> 'SUBJ' -> 2 // ( sind )
.
1 2 'sind'
'cat' / 'VAFIN'
'SYN' -> 'S' -> 0
.
2 3 'begeistert'
'cat' / 'ADJD'
'SYN' -> 'PRED' -> 2 // ( sind )
.
3 4 '!'
'cat' / '$.'
'SYN' -> 0
```

Abbildung 3: Vereinfachte Textdatei im Stil der Hamburg Dependency Treebank

TIGER	TüBa	Beschreibung	Beispiel (in Klammern)
S	SIMPX	Satz	[*Wir sind begeistert*]
AP	ADJX	Adjektivphrase bzw. -chunk	[*noch stärker*]; die [*von seiner Frau geborgten*] *Dollars*
NP	NX	Nominalphrase bzw. -chunk	*von* [*seiner Frau*]
–	VXFIN	finiter Verbalchunk	*Er* [*siegte*]
VP	VXINF	nicht-finite Verbalphrase bzw. -chunk	*Sie will* [$_{VF}$ *es lösen*]; *Sie will es* [$_{VXINF}$ *lösen*]
VZ	–	Infinitiv mit *zu*	*die Wahl* [*zu gewinnen*]
CS	–	koordinierte Sätze	[*Er fordert nicht, er bittet*]
CNP	–	koordinierte Nominalphrasen	*wie* [*Jachten und Villen*]
–	VF	Vorfeld	[*Sie*] *will es lösen*
–	LK	Linke Satzklammer	*Sie* [*will*] *es lösen*
–	MF	Mittelfeld	*Sie will* [*es*] *lösen*
–	VC	Verbkomplex (Rechte Satzkl.)	*Sie will es* [*lösen*]
–	NF	Nachfeld	*Sie wird fordern* [*zu schließen*]

Tabelle 5: Beispiele nicht-terminaler Tags in TIGER und TüBa

relativ *flache* Strukturen annotiert, d.h. Kategorien werden nur angegeben, wenn die Phrasen komplex sind. Bestehen sie nur aus einem Wort, wird kein eigener Phrasenknoten eingefügt. In Abb. 4 entsprechen die weißen Ovale den nicht-terminalen Knoten und die grauen Kästchen den funktionalen Kantentags. TIGER verwendet ca. 50 funktionale Tags, z.B. *HD*=Kopf, *SB*=Subjekt, *PD*=Prädikativ, *NK*=*Noun Kernel*.

Abbildung 5 zeigt einen analogen Baum aus der TüBa-D/Z, die 40 funktionale Tags vorsieht, z.B. *HD*=Kopf, *ON*=Subjekt (wörtl. Objekt, nominativ), *PRED*=Prädikativ.

Der aufmerksame Leser wundert sich vielleicht über die etwas seltsam klingende Terminologie für die Subjektfunktion in der TüBa-D/Z. Die Bezeichnung *Objekt im Nominativ* (ON) ist der Diskussion geschuldet, ob das Deutsche eine sog. konfigurationale Sprache sei und dem Subjekt damit ein Sonderstatus gegenüber den anderen Ergänzungen eines Verbs eingeräumt werden sollte. Die Entwickler des Tagsets haben sich bei der Bezeichnung offensichtlich gegen die Konfigurationalitätsthese entschieden[58].

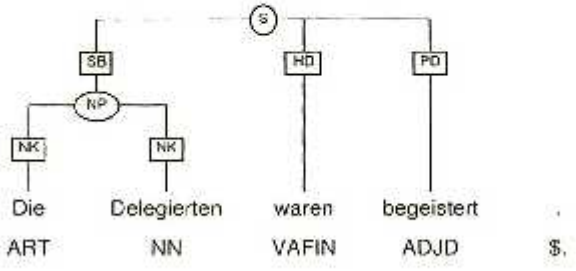


Abbildung 4: Hybride Annotation in der TIGER-Baumbank

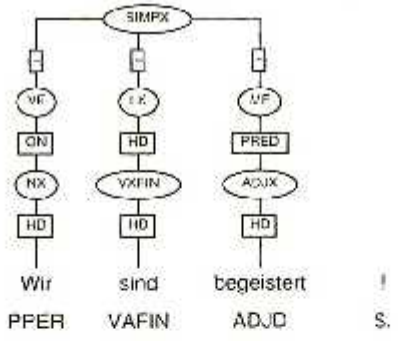


Abbildung 5: Hybride Annotation in den TüBa-Baumbanken

Abbildungen[59] 6 und 7 zeigen etwas komplexere Bäume aus der TIGER-Baumbank. Sie illustrieren zwei Besonderheiten der TIGER-Annotation: *Überkreuzende Kanten*, die bei Stellungsvarianten den syntaktischen Bezug innerhalb eines Satzes festhalten, und *sekundäre Kanten*, die den syntaktischen Bezug bei Koordinationen verdeutlichen, wenn einzelne Konstituenten in einem der Konjunkte fehlen.

---

[58] Vgl. z.B. Haider (1985), für eine Zusammenfassung der Diskussion s. z.B. Fanselow (1987).

[59] An dieser Stelle vielen Dank an Stefanie Dipper, die mehrere Abbildungen dieses Kapitels zur Verfügung stellte.

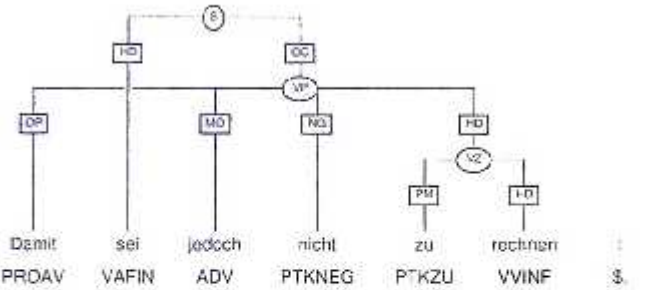


Abbildung 6: Überkreuzende Kanten in TIGER

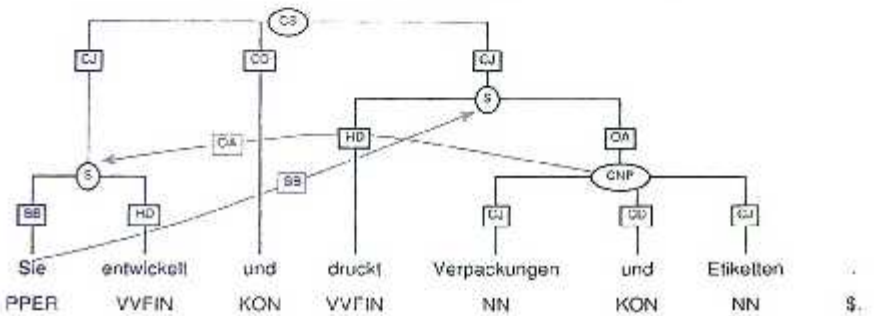


Abbildung 7: Sekundäre Kanten in TIGER

In Abb. 6 ist das topikalisierte Präpositionalobjekt *damit* über eine kreuzende Kante mit dem VP-Knoten der Verbalphrase verbunden. Die sekundären Kanten in Abb. 7 gehören nicht zur eigentlichen Baumstruktur. Sie markieren, dass das Pronomen *Sie* Subjekt sowohl von *entwickelt* als auch von *druckt* ist, und dass die koordinierte NP *Verpackungen und Etiketten* das Objekt beider Verben darstellt.

Wie bei der Dependenzannotation wollen wir Ihnen auch für die phrasenstrukturellen Baumbanken verschiedene Textformate vorstellen und verweisen Sie hierfür auf die Aufgaben am Ende des Kapitels. Sowohl die TIGER-Baumbank als auch die TüBa-Baumbanken wurden mit Hilfe des Annotationswerkzeugs *Annotate* von Thorsten Brants und Oliver Plaehn annotiert[60]. Die Textformate für die Baumbanken sind daher dieselben. Im Aufgabenteil am Ende dieses Kapitels zeigen wir Ihnen den Satz *Wir sind begeistert!* in drei verschiedenen Textformaten.

[60] Vgl. Brants und Plaehn (2000).

4.3.3 Semantische Annotation

Wenn Sie das Stichwort „semantische Annotation" googeln, erhalten Sie wahrscheinlich viele Treffer zum semantischen Web 2.0. Dort werden Webseiten mit Bedeutungskategorien ausgezeichnet, die von Suchmaschinen und anderen Programmen ausgewertet werden können. Darum geht es im folgenden Abschnitt nicht, sondern wir beziehen uns auf eine rein korpuslinguistische Lesart des Terminus.

Semantische Annotationen sind weniger verbreitet als syntaktische Annotationen. Allerdings findet man bereits auf Wortebene semantische Informationen, die als Teil von Wortartentagsets annotiert werden. In TIGER und TüBa-D/Z werden Eigennamen wie *Anna* gemäß STTS mit einem teilweise semantisch motivierten Wortartentag versehen (*NE*). In diesen Baumbanken werden auch auf der Mehrwortebene Namen markiert: Titel wie *„Schlaflos in Seattle"* erhalten ein eigenes, semantisch motiviertes Tag. Ebenfalls auf syntaktischer Ebene werden in der Penn Treebank Adverbiale in Form von Präpositionalphrasen nach ihrer Bedeutung subklassifiziert, z.B. temporale Präpositionalphrasen (*on Friday*) als *PP-temp* oder lokative Präpositionalphrasen (*in Seattle*) als *PP-loc*.

Wiederum auf Wortebene findet die Markierung von einzelnen Lesarten (*Word Senses*) statt. Hierbei werden – meist entlang der Unterscheidung von Lesarten in einer Ontologie oder einem Wörterbuch – verschiedenen Verwendungen einer lexikalischen Einheit unterschiedliche Lesarten zugeordnet (z.B. *Satz*: a. syntaktische Einheit; b. Spielabschnitt im Tennis; c. Menge gleichgeordneter Einheiten; d. großer Sprung usw.). Im amerikanischen OntoNotes-Korpus[61] werden Wortformen in englischen und chinesischen Texten mit semantischen Indizes annotiert, die auf Lesarten-Einträge in der großen, lexikalischen Online-Ressource *WordNet* verweisen[62].

Die Annotation von semantischen Rahmen (*Frames*), im Sinne der *Frame Semantics*[63], geht über die Wortebene hinaus. Ein Frame besteht aus einem Prädikat und allen Argumenten oder Adjunkten, die eine Rolle in Bezug auf das Prädikat spielen. Die Rollen sind verwandt mit den thematischen Rollen der Generativen Grammatik, auch wenn sie weniger allgemein sind[64]. In Beispiel (36) wird das Verb *gilt* als frame-hervorrufendes Element annotiert. Es triggert den Frame *Kategorisierung* („Categorization").

(36) Der Bundestag gilt als Vorbild.

Zwei Rollen des Kategorisierungsframes sind im Satz realisiert: Das *Objekt* („Item") durch die Nominalphrase *der Bundestag* und die *Mitteilung* („Message") durch die Präpositionalphrase (bzw. Adjunktorphrase) *als Vorbild*. Das Saarbrücker SALSA-Projekt annotierte über den syntaktischen Annotationen der TIGER-Baumbank semantische Frames[65].

---

61 Vgl. Hovy et al. (2006).

62 WordNet: `https://wordnet.princeton.edu/`.

63 Vgl. Fillmore (1968).

64 Interessanterweise gehen sowohl das abstrakte Theta-Rollenset der Generativen Grammatik als auch die daten-orientierten *Frame Elements* der Frame Semantics auf Arbeiten von James Fillmore zurück z.B. Fillmore (1968) und Ruppenhofer et al. (2006). Zu Charles Fillmore siehe auch Abschnitt 1.1.

65 Vgl. Erk et al. (2003). SALSA ist angelehnt an das amerikanische FrameNet Projekt, vgl. `framenet.icsi.berkeley.edu/`. Die amerikanische Penn Treebank wird als *Proposition Bank*

Abschließend wollen wir auf die Groningen Meaning Bank (GMB)[66] verweisen, ein relativ junges, englischsprachiges Korpusprojekt aus Groningen, das satzübergreifende semantische und pragmatische Annotationen nach der Diskursrepräsentationstheorie (Discourse Representation Theory, DRT)[67] online bereitstellt. Das Besondere hierbei ist, dass es sich um eine theoretisch wohl motivierte, tiefe semantische Analyse handelt, ganz anders als die sehr ‚flachen' Analysen, die sich oftmals hinter der semantischer Annotation von Korpora verbergen. Allerdings müssen wir hier warnend darauf hinweisen, dass die Annotationen automatisch erstellt wurden und daher sehr fehlerhaft sein können, es sei denn, sie wurden manuell nachkorrigiert. Eine manuelle Korrektur erkennen Sie an sogenannten *Bits of Wisdom*. Geplant ist, dass die Analysen mittels eines Onlinespiels, eines *Games with a Purpose*, nach und nach korrigiert werden. Ob sich dieses Korrekturmodell in der Praxis bewährt, muss sich erst noch zeigen. Genau genommen gehen die Annotationen in der GMB über die klassische, satzbezogene Semantik hinaus und leiten damit über zur pragmatischen Annotation im nächsten Abschnitt.

4.3.4 Pragmatische Annotation

Konzentriert sich die semantische Annotation noch auf die Wort- oder Satzebene, überschreitet man diese Grenzen sehr schnell, wenn man pragmatische Phänomene analysieren will.

Bei der *Anaphern-* oder *Koreferenzannotation* wird eine *Anapher*, z.B. ein Pronomen oder eine definite Nominalphrase, mit einem Bezugswort (*Antezedens*) in Relation gesetzt. Diese Relation markiert, dass der Leser auf die Bedeutung des Antezedens zurückgreifen muss, um die Bedeutung der Anapher im gegebenen Kontext verstehen zu können. Im Falle von Pronomen liegt es z.B. auf der Hand, dass sie alleine nicht genügend Information liefern, um eine Person oder ein Objekt neu im Diskurs zu etablieren (deiktische Pronomen wie *ich* oder *du* sind hier ausgenommen).

Für Leser mit einem Hintergrund in der Generativen Grammatik sei hier erklärend angemerkt, dass sich die Terminologie in der Korpuslinguistik von der Terminologie der sogenannten *Bindungstheorie*[68] unterscheidet, mit der in der Generativen Grammatik satzinterne Koreferenzbeziehungen analysiert werden.

Die Relation zwischen einer Anapher und ihrem Antezedens ist potenziell satzübergreifend und stellt damit besondere Anforderungen an die Annotation und auch an die Korpusabfrage dar. Die Auflösung solcher Koreferenzrelationen ist wichtig, wenn man Informationen in einem Text erschließen möchte. Für Sie als Leser ist es wahrscheinlich trivial, dass in Beispiel (37) mit dem Pronomen *sie* auf die nachgestellte Nominalphrase *die 220 Albaner aus dem Kosovo* Bezug genommen wird. Bei einer automatischen Auswertung ist dieser Bezug nicht ohne Weiteres klar. Wenn z.B. mittels eines Frage-Antwortprogramms die Information gefunden werden soll, wer seit vier Wochen in Berlin ist, dann muss die Anapher mit dem Bezugselement in Relation gesetzt werden. Im

(Palmer et al., 2005) mit semantischen Informationen zu verbalen Argumenten erweitert und ist in das bereits genannte OntoNotes-Korpus integriert.

[66] Groningen Meaning Bank: `http://gmb.let.rug.nl/`.

[67] Vgl. Kamp und Reyle (1993).

[68] Vgl. Chomsky (1981).

engeren Sinn spricht man in diesem Beispiel von einer *Katapher*, da sich das Bezugswort im nachfolgenden Text befindet.

(37) Vier Wochen sind [sie] nun schon in Berlin, [die 220 Albaner aus dem Kosovo].

Korpora werden mit Koreferenzrelationen annotiert, um zu untersuchen, welchen linguistischen Beschränkungen die entstehenden *Referenzketten* unterliegen. Es geht dabei darum, mit welchen Ausdrücken man auf wiederholt erwähnte *Referenten* wie Personen, Objekte und Ereignisse Bezug nehmen kann, so dass der Text kohärent interpretiert wird. Auch für die Entwicklung und das Testen von computerlinguistischen Programmen zur Koreferenzauflösung werden Korpora mit Koreferenzrelationen annotiert[69].

Eine Art Vorstufe zur Koreferenzannotation ist die Annotation mit *Informationsstatus*, der für referierende Ausdrücke angibt, ob deren Referenten bereits vorerwähnt oder dem Hörer anderweitig bekannt sind, oder ob sie neu etabliert werden müssen[70].

Das Stuttgarter DIRNDL-Korpus beinhaltet sowohl Informationsstatus- als auch Koreferenzannotationen[71]. Das Korpus ist dahingehend etwas besonderes, dass es auf vorgelesenen Radionachrichten basiert und als Primärdaten eine textuelle Ebene mit einer Audioebene bzw. den Transkriptionen der Audiodateien verbindet.

Der kleine Text in (38) ist der Anfang einer der Nachrichten[72]. Im ersten Satz wird eine Volksabstimmung in Ägypten eingeführt (*ein Referendum über zahlreiche Verfassungsänderungen*). Im zweiten Satz wird auf dieses Ereignis einmal direkt mit *die Volksabstimmung* und im dritten Satz zweimal indirekt mit den relationalen Ausdrücken *zum Boykott* und *einen fairen Ablauf* Bezug genommen. Der dritte Satz führt außerdem mit *die Opposition* einen zusätzlichen Referenten ein, der zwar nicht vorerwähnt ist, dessen Existenz dem Leser aber grundsätzlich bekannt sein sollte.

(38) [s1] In Ägypten hat ein Referendum über zahlreiche Verfassungsänderungen begonnen. [s2] Allein in der Hauptstadt Kairo sind tausende Polizisten im Einsatz, um die Volksabstimmung abzusichern. [s3] Die Opposition hat zum Boykott aufgerufen, weil sie einen fairen Ablauf nicht gewährleistet sieht.

Tabelle 6 zeigt einen Ausschnitt der Annotationen des dritten Satzes. Das DRNDL-Korpus erfasst zwei Ebenen des Informationsstatus: einen lexikalischen, bezogen auf die reine Wortform, und einen referenziellen, der sich auf die Referenten bezieht, die im Text erwähnt werden[73]. Zum Beispiel ist das Wort *Opposition* im dritten Satz lexikalisch neu, also im Text nicht vorerwähnt (*L-NEW*). Der Referent ist dem Leser aber

---

[69] Siehe z.B. Hinrichs et al. (2004) und Naumann (2005) für die pragmatische Annotation der TüBa-DZ. Ein frei verfügbares Korpus des Englischen ist das *Coreferentially Annotated Corpus*, `clg.wlv.ac.uk/resources/`, Mitkov et al. (2000). Für allgemeine Informationen zur Koreferenzannotation siehe Poesio (2004).

[70] Ein klassische Studie zum Informationsstatus im Englischen stellt Prince (1992) dar. Poesio und Vieira (1998) konzentrieren sich auf definite Nominalphrasen.

[71] DIRNDL: Diskurs-Informations-Radio-Nachrichten-Datenbank für Linguistische Analysen, `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.html`. Das Korpus ist in Björkelund et al. (2014) beschrieben.

[72] Die Satzindizes wurden zum besseren Verständnis hinzugefügt.

[73] Siehe das RefLex-Annotationsschema von Baumann und Riester (2012).

ID	Token	Akzent	Lexikalischer Informationsstatus	Referenzieller Informationsstatus	Koreferenz
1	Die	NONE	-	(R-UNUSED-KNOWN	(712
2	Opposition	L*H	(L-NEW)	R-UNUSED-KNOWN)	712)
3	hat	NONE	-	-	-
4	zum	NONE	-	(R-BRIDGING$2-13-14	-
5	Boykott	H*L	(L-NEW)	R-BRIDGING$2-13-14)	-
6	aufgerufen	L*H	(L-NEW)	-	-
7	,	N/A	-	-	-
8	weil	H*L	-	-	-
9	sie	NONE	-	(R-GIVEN)	-712
10	einen	NONE	-	(R-BRIDGING$2-13-14	-
11	fairen	H*L	(L-NEW)	-	-
12	Ablauf	L*H	(L-NEW)	R-BRIDGING$2-13-14)	-
13	nicht	H*	-	-	-
14	gewährleistet	!H*L	(L-NEW)	-	-
15	sieht	NONE	(L-NEW)	-	
16	.	N/A	-	-	-

Tabelle 6: Annotation von pragmatischen und prosodischen Merkmalen im DIRNDL-Korpus (vereinfachte Darstellung)

trotzdem bekannt, da es Teil des Weltwissens ist, dass in einem Staat eine Opposition existiert (*R-UNUSED-KNOWN*). In der Spalte *Koreferenz* ganz rechts sind Wortgruppen, die einen Referenten benennen, der mehrfach im Text wiederaufgegriffen wird, durch numerische Indizes markiert. Die Indizes sind quasi Namen für die Referenten. Der Referent 712 tritt im dritten Satz zweimal in Erscheinung: *Die Opposition* und *sie* sind koreferent, da sie beide auf den Referenten 712 referieren. Bei Bezügen ohne unmittelbare Koreferenz (*Bridging*) wird der Bezugsausdruck als Code angegeben: Zum Beispiel sind *zum Boykott* und *einen fairen Ablauf* jeweils mit *R-BRIDGING$2-13-14* markiert, da sie sich beide, wie oben bereits erwähnt, indirekt auf den Referenten von *die Volksabstimmung* beziehen (R-BRIDGING), welches durch das 13. und 14. Wort im dem zweiten Satz gebildet wird ($2-13-14).

Das DIRNDL-Korpus wurde mit dem Ziel erstellt, Korrelationen zwischen Informationsstatus und Prosodie untersuchen zu können. Man möchte herausfinden, ob der Satzakzent Informationen über die Bekanntheit oder Neuheit eines Referenten vermittelt. Die tabellarische Darstellung in Tab. 6 beinhaltet daher auch die Annotation der Satzakzente durch Töne (*High* und *Low*) nach dem ToBI-Schema[74].

Ein weiterer Typ von pragmatischer Annotation ist die Informationsstruktur im Sinne von *Topik* (das, wovon der Satz handelt) und *Fokus* (neue Information). Das *Potsdam Commentary Corpus*[75] ist ein Beispiel für diese Art von Annotation. Der Schwerpunkt des

[74] ToBI: Tones and Break Indices, vgl. Silverman et al. (1992).
[75] Vgl. Stede (2004).

PCC liegt allerdings auf einer anderen Art der satzübergreifenden Analyse: Das Korpus wird mit Diskursstrukturen nach der *Rhetorical Structure Theory*[76] angereichert. Dabei werden Sätze und größere Bestandteile des Textes in Bezug zu einander gesetzt, vgl. Abb. 8: Die Diskursrelation *Evaluation* verbindet einen Kommentar mit der kommentierten Situation. Eine *Elaboration* gibt zusätzliche Information zur Kernaussage und eine *Antithesis* zeigt einen Widerspruch auf.

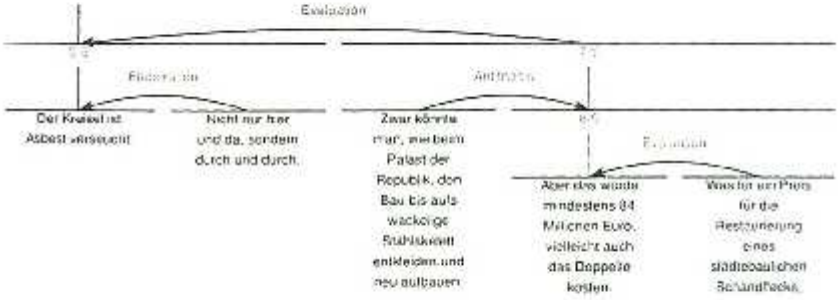


Abbildung 8: Annotation von Diskursstruktur nach der *Rhetorical Structure Theory* im *Potsdam Commentary Corpus* (PCC)

4.4 Normalisierung und Fehlerannotation

Nachdem sich die letzten Abschnitte an den Kerndisziplinen der Linguistik orientiert haben, betrachten wir nun noch Annotationstypen, die darüber hinausgehen.

Texte, die in sich inkonsistent oder fehlerhaft sind, werden oftmals *normalisiert* bevor sie korpuslinguistisch weiterverarbeitet werden. In Texten früherer Sprachstufen, zum Beispiel dem Mittel- oder Frühneuhochdeutschen, findet man oft mehrere Schreibweisen von ein und demselben Lexem (z.B. *sein* vs. *seyn*). Es gab damals keine einheitliche Rechtschreibung, sondern dialektal geprägte Varianten, die in sich ebenfalls nicht einheitlich waren. Auch die einzelnen Schreiber waren für sich genommen nicht unbedingt konsistent, so dass es auch Rechtschreibvariation innerhalb einzelner Texte gibt. Bei der Normalisierung werden die Wortformen auf eine Normschreibung abgebildet, sodass es möglich wird, systematisch zu suchen oder weiterführende Annotationen wie Wortartentagging automatisch durchzuführen. Eine Normalisierung ist ein wenig vergleichbar mit der Grundformzuweisung bei der Lemmatisierung. Dabei kann sich die Norm z.B. auf ein historisches Referenzwörterbuch beziehen wie den Lexer für Mittelhochdeutsch[77] oder man übersetzt ins moderne Standarddeutsch. Analoges gilt für moderne, inkonsistent geschriebene Texte wie z.B. computervermittelte Kommunikation in Chatbeiträgen oder SMSen oder bei Transkriptionen gesprochener Sprache. Die

[76] Vgl. Mann und Thompson (1988), siehe `www.sfu.ca/rst`.
[77] Lexer: `http://woerterbuchnetz.de/Lexer`.

Normalisierung kann eine zusätzliche Annotationsebene bilden, die genauso durchsucht werden kann wie die Ebene der Wortarten- oder Lemmaannotationen. Wichtig ist, dass der Bezug zum Originaltext immer bestehen bleibt, weil man sonst die Gefahr läuft, interessante Muster und Entwicklungen zu übersehen, falls sie in der „Übersetzung" nicht abgebildet werden. Das gilt besonders für historische Sprachstufen, die in ihrer Lexik und Grammatik vom heutigen Standarddeutsch abweichen[78].

Ein weiterer Korpustyp, bei dem Normalisierung eine große Rolle spielt, ist das Lernerkorpus. Ein Lernerkorpus enthält typischerweise Texte oder Transkripte von Lernern einer Fremdsprache. Selten wird der Terminus auf Korpora zum Erstspracherwerb oder Schriftspracherwerb angewendet[79]. Bei der Untersuchung von Lernersprache ist man an echten Abweichungen von der Zielsprache („Fehlern"), aber auch an der Lernersprache ansich als Interimssprache bzw. *Interlanguage* interessiert. Die Korpora dienen als Grundlage für computergestützte Analysen, die sowohl für den Fremdsprachunterricht relevant sein können als auch für die Fremdspracherwerbsforschung[80]: Bei der *Contrastive Interlanguage Analysis* werden systematische Abweichungen der Lernersprache von einer Kontrollvariätet untersucht, zum Beispiel in Bezug auf die Auftretenshäufigkeiten von bestimmten Wörtern oder Konstruktionen, die für sich genommen durchaus grammatisch sein können. Die Methode der computerunterstützten Fehleranalyse sieht darüber hinaus vor, dass im Korpus lernersprachliche Abweichungen von der Zielsprache markiert werden. Diese Annotation setzt voraus, dass man eine Vorstellung von der jeweiligen normhaften Ausprägung besitzt, d.h. eine sogenannte *Zielhypothese* formuliert.

Abweichungen werden auf allen Ebenen der Sprache beobachtet: bei der Aussprache, in der Orthographie, der Morphologie, bei der Wahl von Tempus oder Modus, bei der Kongruenz zwischen Wortformen, bei der Wortstellung usw. Es können auch Angemessenheitsfehler vorkommen, wie z.B. dass eine idiomatische Wendung falsch eingesetzt wird.

Das Berliner Lernerkorpus Falko ist eine Sammlung von linguistisch aufbereiteten und fehlerannotierten Lernertexten sowie muttersprachlichen Vergleichstexten[81]. Anstelle eines feinkörnigen Fehlertagsets beinhalte: Falko zwei Normalisierungsebenen: *Zielhypothese 1* (ZH1), mit minimalen, rein satzbezogenen, grammatischen Korrekturen und *Zielhypothese 2* (ZH2), die auch semantische und pragmatische Korrekturen beinhaltet, so dass die einzelnen Sätze nicht nur für sich genommen, sondern auch im Textzusammenhang sinnvoll und kohärent erscheinen[82].

Das Konzept Zielhypothese darf hier nicht als Rekonstruktion der Lernerintention missverstanden werden. Was der Lerner im Moment des Schreibens wirklich ausdrücken

---

78 Krasselt et al. (2015) beschreiben detailliert die Normalisierung des frühneuhochdeutschen Anselm-Korpus, das auf standarddeutsche Formen abgebildet wird.

79 Ein Beispiel eines Lernerkorpus von muttersprachlichen Lernern ist z.B. das KoKo-Korpus das unter der Leitung von Andrea Abel an der Europäische Akademie Bozen (EURAC) aufgebaut wird, vgl. `http://www.korpus-suedtirol.it/`.

80 Vgl. Nesselhauf (2004); Lüdeling und Walter (2010).

81 Vgl `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/Falko/standardseite`.

82 Die Falko-Guidelines sind in Reznicek et al. (2012) beschrieben; Reznicek et al. (2013) erläutern das Konzept der Zielhypothesen in Falko.

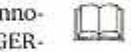
wollte, ist leider nicht rekonstruierbar. Es sei denn, man hätte so etwas wie einen Mitschnitt seiner Gedanken. Etwas leichter ist es, wenn der Lernertext als Nacherzählung einer Geschichte (auch einer Bildgeschichte oder eines Films) entstanden ist. Aber auch hier hat man schlussendlich keine Gewissheit.

Die Zielhypothesen im Falko-Korpus stellen Normalisierungsebenen dar, deren Erstellung durch Regeln in den Guidelines festlegt ist. Ein automatischer Abgleich der Originalsätze mit einer der Zielhypothesen erzeugt an den Stellen, an denen das Original von der Norm abweicht, *Editiertags*, welche die Veränderungen auf der Wortebene mechanisch dokumentieren: *INSert*, *DELete*, *CHAnge*, *SPLIT*, *MOVeSource* und *MOVeTarget*. Dieser Abgleich kann auf der Textebene und auch auf allen Annotationsebenen durchgeführt werden[83]. Für eine weiterführende Fehleranalyse ist es meistens notwendig, die Editiertags weiter zu interpretieren. Anke Lüdeling und ihre Kollegen vom Falko-Korpus[84] argumentieren dafür, einzelne Fehlertypen getrennt voneinander zu annotieren (z.B. Genus-Fehler getrennt von Numerusfehlern). Dies ist hilfreich, wenn ein Wort oder eine Sequenz gleichzeitig von verschiedenen Fehlertypen betroffen ist (z.B. gleichzeitig von einem Kongruenz- und einem Kollokationsfehler). Eine fehlerspezifische Annotation erlaubt es auch, alternative Zielhypothesen zu formulieren, wenn eine Abweichung auf mehrere Art und Weise erklärt werden kann. Bei einem Kongruenzfehler wie in Tabelle 7 ist es z.B. nicht immer klar, welches Wort tatsächlich falsch gebildet ist[85]: Besteht der Fehler darin, dass das attributiv verwendete Pronomen *diese* fälschlicherweise als Femininum flektiert wurde, oder steht das Nomen *Phänomen* im falschen Numerus?

Token	Die	Erklärung	für	diese	Phänomen	...
ZH_{Genus}				dieses		
Fehler				Genus		
$ZH_{Numerus}$					Phänomene	
Fehler					Numerus	

Tabelle 7: Annotation konkurrierender Fehleranalysen in Falko

4.5 Weiterführende Literatur



Ule und Hinrichs (2004) geben weiterführende Informationen zur linguistischen Annotation im Allgemeinen und einen Vergleich der Annotationsschemata von der TIGER- und den TüBa-D/Z-Baumbanken im Besonderen. Eine umfassende Übersicht über internationale Baumbankprojekte finden Sie bei Nivre (2008). Rehbein (2010) fasst den State-of-the-Art von dependenzannotierten Korpora zusammen. Annotationen auf allen linguistischen Ebenen und deren korpuslinguistische Nutzung stehen im Zentrum

[83] Reznicek und Zinsmeister (2013) diskutieren z.B. den Abgleich von Original und Zielhypothese in Bezug auf die Annotation mit Wortarten.

[84] Vgl. Lüdeling et al. (2005b); Reznicek et al. (2013).

[85] Vielen Dank an Maik Walter, der die Tabelle zur Verfügung stellte.

der englischsprachige Einführung in die Korpuslinguistik von Kübler und Zinsmeister (2015).

Wie auch in anderen Bereichen dieses Buches haben wir die Annotation von gesprochener Sprache (phonetische und prosodische Annotation) sowie multi-modale Annotation ausgeklammert[56]. Eine ausführliche Übersicht zur Literatur und zu Werkzeugen der phonetischen und prosodischen Annotation finden Sie auf der Webseite von EXMARaLDA (`www.exmaralda.org`). Eine allgemeine Einführung in das Thema *Sprachdatenbanken* bietet Draxler (2008). Der Sammelband von Schmidt und Wörner (2012) stellt nicht nur multi-linguale, sondern auch Korpora gesprochener Sprache und multi-modale Korpusvorhaben vor.

4.6 Aufgaben

1. Warum werden Korpora annotiert? Erklären Sie, warum Sprachwissenschaftler und Computerlinguisten die zeitaufwändige Aufgabe der Annotation auf sich nehmen.
2. Uns fällt es meistens überhaupt nicht auf, wie ambig Sprache ist. In Witzen wird das manchmal ausgenutzt.

   A: Wer kann mir sagen, wie lange Europäer im Durchschnitt studieren?
   B: Genauso wie kurze.

   Analysieren Sie die Äußerung von Sprecher A. Unterscheiden Sie die beiden Lesarten, indem Sie den Satz mit STTS-Tags annotieren, vgl. Tab. 3 auf S. 66, und die Konstituentenstruktur durch Klammerung skizzieren. Welche Tags erhält *lange*?
3. Zeichnen Sie eine Baumstruktur für die gechunkte Nominalgruppe in Beispiel (35) auf S. 76 (das wir hier als Beispiel (39) nochmals wiederholen). Welche der Knoten sind rekursiv?

   (39) [$_{NC}$ die [$_{AC}$ [$_{PC}$ durch [$_{NC}$ Fehlentscheidungen]] | [$_{AC}$ hochverschuldete]] Bahn]
4. Diese letzte Aufgabe richtet sich an die „Tüftler" unter den Lesern. Die drei auf der nächsten Seite folgenden Abbildungen sind drei alternative Repräsentationen des Satzes *Wir sind begeistert!*, vgl. auch Abb. 5 auf S. 79. Bei den Repräsentationsformaten handelt es sich um:

   - Indizierte Klammerstruktur (*Labeled Bracketing Format*) – auch als „Penn-Treebank-Stil" bezeichnet
   - Spaltenformat, auch „(NEGRA-)Export-Format" genannt
   - XML-Repräsentation

   Ihre Aufgabe ist es, jeweils nachzuvollziehen, wie die Wörter des Satzes und die Annotationen in den drei Formaten kodiert werden. Am besten gehen Sie anhand der drei folgenden Leitfragen vor:

   a) Wie werden terminale und nicht-terminale Knoten dargestellt?
   b) Wie werden die Kanten des Baumes – also die Verbindungslinien – kodiert?
   c) Wo findet man die funktionalen Kantentags?

[56] Die einzige Ausnahme ist der Verweis auf die prosodische Annotation im DRINDI-Korpus, siehe Abschnitt 4.3.4.

```
%%Satz 1630
(   (SIMPX
      (VF
         (NX-ON
            (PPER-HD Wir)))
      (LK
         (VXFIN-HD
            (VAFIN-HD sind)))
      (MF
         (ADJX-PRED
            (ADJD-HD begeistert))))
   ($. !))
```

Abbildung 9: Klammerstruktur der TüBa-D/Z

```
Wir                 PPER    --          HD      500
sind                VAFIN   --          HD      501
begeistert          ADJD    --          HD      502
!                   $.      --          --      0
#500                NX      --          ON      503
#501                VXFIN   --          HD      504
#502                ADJX    -           PRED    505
#503                VF      --          -       506
#504                LK      --          -       506
#505                MF      --          -       506
#506                SIMPX   --          --      0
```

Abbildung 10: Spaltenformat der TüBa-D/Z

```
<sentence>
   <node cat="SIMPX" func="--" parent="0" comment="">
     <node cat="VF" func="-" comment="">
       <node cat="NX" func="ON" comment="">
         <word form="Wir" pos="PPER" func="HD" comment=""/>
       </node>
     </node>
     <node cat="LK" func="-" comment="">
       <node cat="VXFIN" func="HD" comment="">
         <word form="sind" pos="VAFIN" func="HD" comment=""/>
       </node>
     </node>
     <node cat="MF" func="-" comment="">
       <node cat="ADJX" func="PRED" comment="">
         <word form="begeistert" pos="ADJD" func="HD" comment=""/>
       </node>
     </node>
   </node>
   <word form="!" pos="$." func="--" parent="0" comment=""/>
</sentence>
```

Abbildung 11: XML-Format der TüBa-D/Z

5 Übung macht den Meister — Annotation im praktischen Einsatz

Wenn man die Einheitskost nicht mag, kocht – sprich: annotiert man selber. In diesem Kapitel soll der Weg zur eigenen Annotation aufgezeigt werden.
Am Ende dieses Kapitels wissen Sie, mit welchen Werkzeugen Sie annotierte Korpora sichten und durchsuchen können. Sie haben Standards und Methoden für das eigene Annotieren kennengelernt und dabei neben verschiedenen Annotationstools auch die Methode des Annotationszyklus gesehen, die wir Ihnen für die Datenanalyse ganz allgemein ans Herz legen wollen.

5.1 Suche in Korpora

5.1.1 Online-Schnittstellen

Nachdem Sie nun einen gewissen Überblick über Korpora und linguistische Annotationsebenen erhalten haben, widmen wir uns nun der Frage, wie Sie auf Texte und deren Annotationen zugreifen können.

Viele Korpusprojekte bieten Online-Abfragemöglichkeiten ihrer Korpora an. Vorreiter in der deutschen Korpuslinguistik ist das Institut für Deutsche Sprache in Mannheim, das mit COSMAS[1] seit Jahren externe Abfragen auf der Mannheimer Korpussammlung ermöglicht. Der Zugang ist kostenlos, man muss sich lediglich als Nutzer registrieren. Das wortarten-getaggte Kernkorpus des DWDS-Projekts in Berlin kann man auch ohne vorherige Anmeldung abfragen. Die kostenlose Registrierung ist trotzdem empfehlenswert, da sie den Zugriff auf eine größere Datenmenge freigibt[2]. Über das in Dänemark angesiedelte *Visual Interactive Syntax Learning*-Projekt (VISL-Projekt)[3] hat man mit *CorpusEye* einen nutzerfreundlichen Online-Zugriff auf Korpora mit Dependenzstrukturen[4]. Abschließend wollen wir noch auf das *Open Source Parallel Corpus*-Portal (OPUS)[5] verweisen, über das man online auf diverse Textsorten mit Übersetzungen in viele Sprache zugreifen kann. Genau genommen handelt es sich um wortarten-getaggte Texte und

---

1 COSMAS: `www.ids-mannheim.de/cosmas2/`.
2 DWDS: `www.dwds.de`.
3 VISL-Projekt: `corp.hum.sdu.dk/cqp.de.html`.
4 Genau genommen sind die VISL-Korpora mit kategorialgrammatischen Analysen angereichert (Karlsson, 1990), welche wiederum eine Dependenzstruktur zugrundelegen.
5 OPUS: `http://opus.lingfil.uu.se/`.

deren Übersetzungsäquivalente. Eine immer größer werdende Anzahl von Korpora, einschließlich der beiden Lernerkorpora Falko und Kobalt, können Sie über das Online-Suchtool *ANNIS* abfragen[6].

5.1.2 Suchwerkzeuge und Abfragesprachen

Idealerweise wird ein Korpus von einem Suchwerkzeug (auf Englisch ‚Query Tool') begleitet. Ist das nicht der Fall, kann man auf eine Reihe von kostenlosen Werkzeugen zurückgreifen, die man sich aus dem Internet herunterladen kann. Wir geben hier nur ein paar Anregungen und verweisen wieder auf die Webseite zum Buch, auf der Sie weitere Informationen finden können.

Die einfachste Suche läuft über die Wortformen. Insbesondere für lexikografische Fragestellungen und im Bereich des Sprachenlernens kann man hier wertvolle Informationen finden. Als Darstellungsform eignet sich eine *Konkordanz* (auch *Keyword in Context*, KWIC-Format), die die einzelnen Treffer untereinander auflistet und jeweils einen gewissen Ausschnitt aus dem vorangehenden und folgenden Text ausgibt. Abbildung 12 zeigt den Ausschnitt einer Suche von *begeistert* auf dem IMS-Korpus der *Frankfurter Rundschau* mit der *Corpus Workbench* des Instituts für Maschinelle Sprachverarbeitung in Stuttgart (IMS). Anhand der Konkordanz kann man z.B. Hypothesen über die verschiedenen Lesarten eines Wortes bilden. In Abb. 12 sind Belege für die Verwendungsweisen von *begeistert* aufgelistet. Man kann die Adjektivlesart von der Verblesart (*sich begeistern*) unterscheiden und findet auch zwei Steigerungspartikel (*wenig*, *total*).

```
sie auf den Wahlversammlungen  begeistert zujubeln . Anderen gilt er al
gewesen , erzählt er sichtlich begeistert , und es gebe wohl keinen oet
detaillierte Beschreibung \ ,  begeistert sich Scheftel: \ Shaleyev ka
halk ist von dieser Idee wenig begeistert . Den Umbau , der nötig wäre
ildet . \ Die Leute sind total begeistert von dieser Idee \ , freut sic
```

Abbildung 12: Konkordanz (*KWIC*-Format, *KeyWord in Context*)

Bei der linguistischen Suche entsteht schnell das Bedürfnis nach einer ausdrucksstärkeren Suchmöglichkeit als der Suche nach Wortvollformen. Man möchte Anfragen unterspezifizieren, weil man z.B. gleichzeitig nach verschiedenen Flexionsformen einer Grundform suchen möchte oder man ist an Wortgruppen interessiert, von denen man aber nur einen Teil spezifisch vorgeben kann oder will. Kurz gesagt, man möchte nicht nach einzelnen Wortformen, sondern nach Mustern im Text suchen. Dies kann man über *reguläre Ausdrücke* erreichen. Der folgende Exkurs stellt eine standardisierte Variante von regulären Ausdrücken vor. Je nach Tool haben Sie es aber mit unterschiedlichen Varianten von Ausdrücken zu tun.

[6] ANNIS: `http://annis-tools.org/`.

Exkurs: Reguläre Ausdrücke

Reguläre Ausdrücke bzw. Platzhalterzeichen und Operatoren sind Ihnen möglicherweise schon durch die Bedienung von Suchmaschinen bekannt. Suchmaschinen bieten zumindest in der Expertensuche die Anwendung von Platzhalterzeichen an, wenn diese auch nicht immer so vollständig sind wie die hier vorgestellte reguläre Sprache. Ein regulärer Ausdruck beschreibt ein bestimmtes Textmuster in einer abgekürzten oder unterspezifizierten Form. Dazu sind eine Reihe von *Metazeichen* definiert.

Jeder, der schon einmal einen Ausdruck oder ein linguistisches Phänomen in einem Korpus gesucht hat, weiß, wie praktisch es ist, wenn man nach mehreren Wortformen gleichzeitig suchen kann. Ein ganz einfaches Beispiel ist die Suche nach den alternativen Wortformen *Rad* und *Rat*. Der *oder*-Operator ‚|' trennt die beiden Alternativen.

(1) *Alternation* (*oder*-Verknüpfung): |
`(Rad|Rat)` → findet alle Vorkommen von *Rad* und *Rat*.

Die runden Klammern markieren den Bezugsbereich der Alternation. Bei längeren Wörtern kann man sich Tipperei ersparen, wenn man die Alternation auf den gemeinsamen Wortteil beschränkt.

(2) *Gruppierung*: ( )
`Ra(d|t)` → findet ebenfalls alle Vorkommen von *Rad* und *Rat*, vgl. Beispiel (1).

Wenn Sie alle Wörter suchen wollen, die mit *Ra* beginnen und insgesamt drei Buchstaben haben, verwenden Sie für den dritten Buchstaben einen Platzhalter. Der Platzhalter wird oft durch einen einfachen Punkt dargestellt.

(3) *Platzhalter* (*wildcard*): .
`Ra.` → findet z.B. *Rad, Ray, Rat, Rap, Rau* und *Ram*. Der Platzhalter steht für genau ein weiteres Zeichen.

Wenn man anstelle des Platzhalters nur bestimmte Zeichentypen zulassen möchte, kann man eine *Zeichenklasse* festlegen. Anstatt des Punktes verwendet man dann eckige Klammern und listet alle Zeichen auf, die zugelassen werden sollen.

(4) *Zeichenklasse*: [ ]
`Ra[dt]` → findet alle Vorkommen von *Rad* und *Rat*, vgl. Beispiele (1) und (2).

Vielleicht fragen Sie sich jetzt, wie sich eine Suche mit Alternation von einer Suche mit Zeichenklasse unterscheidet. Eine Alternation kann mehrere Zeichen umfassen. Eine Zeichenklasse listet die Alternativen für genau *ein* Zeichen auf[7].

(5) `Ra(d|t|um)` → findet *Rad, Rat* und *Raum*.
`Ra[dtum]` → findet *Rad, Rat* und *Rau, Ram*.

[7] Der Vorteil der Suche über Zeichenklassen ist, dass sie normalerweise vom Computer schneller verarbeitet werden können.

Sehr praktisch ist eine negative Suche, bei der man bestimmte Zeichen explizit ausschließen kann. Ein Möglichkeit dafür bietet die Suche über eine *negierte Zeichenklasse*. Bitte beachten Sie, dass man auf diese Weise nur einzelne Zeichen negieren kann, nicht ein ganzes Wort.

(6) *Negierte Zeichenklasse* [^]:
`[^R]at` → findet alle dreistelligen Wörter, die mit *at* enden, aber nicht mit *R* anfangen: *bat, Cat, hat, Hat, Pat, Sat, tat, Tat*[8] usw., aber auch *rat*, weil nur das Zeichen *R* ausgeschlossen ist, nicht das Zeichen *r*.

Manchmal möchte man die Suche unterspezifizieren und nur einen Teil des Wortes festlegen. Den nicht festgelegten Teil kann man z.B. durch wiederholtes Aufrufen des Platzhalters abdecken. Für Zeichenwiederholungen verwendet man Operatoren[9].

(7) *Operator für optionales Auftreten* (= kein- oder einmal): ?
`Rat.?` → findet *Rat*, *Rats* und zum Beispiel auch den Eigennamen *Rath*.

(8) *Operator für ein- oder mehrfaches Auftreten*: +
`Rat.+` → findet *Rats, Rates, Raten, Rathaus, Rathausmarkt, Ratlosigkeit*, usw., aber nicht *Rat*, weil der Operator verlangt, dass der Platzhalter mindestens einmal durch ein Zeichen ersetzt wird.

(9) *Operator für kein- oder einmaliges oder beliebig häufiges Auftreten*: * (auch *Kleene*-Stern genannt)
`Rat.*` → findet dieselben Vorkommen wie (8) plus zusätzlich auch *Rat*, weil der Operator auch Optionalität zulässt.
`.*[rR]at.*` → geht noch einen Schritt weiter. Es findet alle Wörter, die irgendwo im Wort (auch am Anfang oder Ende) die Sequenz *rat* oder *Rat* haben. Also auch *beraten, Bundesrat* oder *Bundes-Rat*, aber auch *Demokratie* oder *Strategie*.

Bei den bisher genannten Beispielen sind wir immer davon ausgegangen, dass Anfang und Ende des Suchmusters auch Anfang und Ende eines Wortes im Text beschreiben würden. In den Suchtools *TIGERSearch* und *CQP* ist das tatsächlich so (bzw. die Grundeinstellung). In anderen Suchtools müssen Sie diese Grenze evtl. explizit markieren[10].

Vielleicht meinen Sie jetzt, dass es ja schön und gut sei mit den regulären Ausdrücken und ihren Metazeichen, aber dass Sie eigentlich an Abkürzungs- und Satzendepunkten interessiert seien. Wie kann man danach suchen, wenn der Punkt als Platzhalter doch für jedes beliebige Zeichen stehen kann? Die Lösung ist einfach. Alle Metazeichen verlieren durch einen vorangestellten *Backslash* (auch *Rückstrich*) „\" ihre besondere Bedeutung. Die Folge *Backslash-Punkt* („\.") steht z.B. für das Punktzeichen.

---

[8] *Cat* geht auf den Ausdruck *Cat Eye* im Korpus zurück. *Hat* ist ein großgeschriebenes *hat* am Satzanfang. *Pat* stammt von *Pat Lewis*, *Sat* schließlich von *Sat 1*.

[9] Operatoren können in Bezug auf alle Zeichen oder Gruppierungen verwendet werden, nicht nur zusammen mit dem Platzhalterzeichen. Die Suche nach `12(34)*` findet z.B. *12, 1234, 123434* oder auch *1234343434*.

[10] Die Bezugnahme auf Wortanfang und -ende (oder Zeilenanfang und -ende) nennt man auch *Wortanker* (bzw. *Zeilenanker*.)

Die hier vorgestellten Metazeichen und ihre Bedeutungen stellen einen gewissen Standard dar, werden aber nicht in allen Anwendungen genauso verwendet. Es ist daher immer wichtig, vor der Verwendung einer Abfragesprache die dazugehörige Dokumentation zu lesen.

Fortsetzung: Suchwerkzeuge und Abfragesprachen

Ein ausdrucksstarkes Suchtool bietet die *IMS Open Corpus Workbench* (CWB) die ursprünglich am Institut für Maschinelle Sprachverarbeitung in Stuttgart entwickelt wurde[11]. Die zugrundeliegende Abfragesprache heißt *CQP* (für *Corpus Query Processor*)[12]. CWB wird weltweit von einer ganzen Reihe von Online-Korpusprojekten für die Abfrage eingesetzt. Über eine Online-Demo der CWB können Sie ein Korpus mit Bundestagsdebatten sowie Ausschnitte aus dem Parallelkorpus *Europarl* durchsuchen. Beide Korpora sind mit Wortarteninformationen (für das Deutsche jeweils mit dem STTS-Tagset) und partieller Satzanalyse (Chunking) annotiert[13]. Sie können sich die Corpus Workbench *CWB* auch auf dem eigenen Rechner installieren. Das Tool umfasst neben dem eigentlichen Suchprogramm auch ein Programm, mit dem Sie neue Korpora ins CWB-Format überführen und ins Tool einlesen können[14].

Wesentlich einfacher zu installieren ist das Korpustool *AntConc*[15] von Laurence Anthony, einem englischen Korpuslinguisten, der in Japan arbeitet. Das Tool ist sehr gut beschrieben einschließlich einer Reihe von Videotutorials. Ähnlich wie CWB kann man mit AntConc wortarten-annotierte Korpora auswerten und eine Variante des Tools, *AntPConc*[16] kann auch für die Suche auf parallelen Korpora eingesetzt werden. Darüber hinaus unterstützt AntConc das Auffinden von Kollokationen im Korpus.

Ein ähnliches Programm zur Darstellung und Suche von Wortformen und Kollokationen in Texten ist das kostenpflichtige *WordSmith*[17] von Mike Scott. Sie können sich dieses Tool auch als freie Demo-Version herunterladen und haben dann Zugriff auf immerhin 50 Treffer pro Anfrage.

Ein anderes Suchtool wurde im Rahmen des *TIGER* Projekts speziell für die Suche auf (Konstituenten-)Baumstrukturen entwickelt: *TIGERSearch*[18]. Zusätzlich zu einer CQP-ähnlichen textbasierten Suche bietet es auch die Option eines grafischen Suchinterfaces an. Diese Option eignet sich ganz speziell für Linguisten ohne Vorkenntnisse in Abfragesprachen. Man kann sich (Teil-)Strukturen „zusammenklicken", aber auch reguläre Ausdrücke integrieren. Das Tool generiert auf Abruf die entsprechende textuelle Suchanfrage, so dass man nach und nach die Abfragesyntax lernen kann[19]. Auch wenn

---

11 IMS Corpus Workbench: `http://cwb.sourceforge.net/`.

12 Vgl. Christ und Schulze (1995) und Evert (2010).

13 Die CQP-Demokorpora sind mit Stefan Evert an die Universität Erlangen umgezogen: `http://corpora.linguistik.uni-erlangen.de/demos/CQP/cqpdemo.html`.

14 CWB wurde ursprünglich für die Betriebssysteme SUN Solaris und Linux geschrieben. Über Portierungs-Software wie CYGWIN `de.wikipedia.org/wiki/Cygwin` kann es auch auf Windows-Rechnern eingerichtet werden.

15 AntConc: `http://www.laurenceanthony.net/software/antconc/`.

16 AntPConc: `http://www.laurenceanthony.net/software/antpconc/`.

17 WordSmith: `http://www.lexically.net/wordsmith/`.

18 Vgl. Lezius (2002).

19 Auf S. 98 finden Sie ein Beispiel für eine grafische Anfrage.

das Tool im Rahmen des TIGER-Projekts entwickelt wurde, ist es nicht auf die TIGER-Baumbank beschränkt. Es ist mit einem Konversionstool gekoppelt (*TIGERRegistry*), das Filter für verschiedene gängige Korpusformate anbietet, die auf diese Weise in TIGERSearch integriert werden können. TIGERSearch ist für Forschungsvorhaben kostenlos. Leider konnte es in den letzten Jahren nicht weiterentwickelt werden, sodass Sie ggf. Probleme haben werden, das Tool auf neueren Betriebssystemen zu installieren. Sollte dies der Fall sein, wollen wir Ihnen zwei neuere Tools empfehlen, die viele Eigenschaften von TIGERSearch aufgegriffen haben: TüNDRA und ANNIS.

TüNDRA ist ein rein web-basiertes Tool. Es wurde im Rahmen des CLARIN-Projekts entwickelt und ist für akademische Nutzer nach Anmeldung frei zugänglich[20]. Sie haben dort Zugriff auf die Tübinger Baumbanken (u.a. TüBa-D/Z im Konstituenten-Format sowie einer automatisch erzeugten Dependenz-Version) und ein paar weitere Korpora. Außerdem können Sie dort eigene, syntaktisch annotierte Korpora, die Sie z.B. mit den WebLicht-Tools annotiert haben (siehe Abschnitt 5.4), hochladen und anschließend durchsuchen.

Das zweite hier zu nennende Nachfolgetool, *ANNIS*[21], haben wir bereits in Abschnitt 5.1.1 im Zusammenhang mit dem Online-Zugriff auf Korpora erwähnt. Es deckt nicht nur syntaktische Strukturen ab, sondern ist sehr allgemein gehalten. Es ist ein sog. generischen Suchtool für die Suche auf komplex annotierten Korpora, einschließlich der Darstellung von multimodalen Inhalten wie Videosequenzen. ANNIS wurde im Rahmen des Sonderforschungsbereichs *Informationsstruktur* der Universität Potsdam und der Humboldt-Universität zu Berlin entwickelt. Einen ersten Eindruck der Möglichkeiten von ANNIS erhalten Sie über das *Tutorial* der Online-Demoversion[22]. Ähnlich wie TIGERSearch erlaubt ANNIS, Suchanfragen auch über eine grafische Schnittstelle zusammenzustellen (siehe den Button ‚Query Builder'), was für Anfänger eine echte Hilfe darstellen kann. Neben der Webinstallation können Sie ANNIS auch lokal auf Ihrem eigenen Rechner installieren, vorausgesetzt, dass Sie das Datenbank-Managementsystem *PostgreSQL* vorinstalliert haben. Letzteres sollte keine Hürde darstellen, da PostgreSQL für die gängigen Betriebssysteme kostenfrei zur Verfügung steht[23]. Mittels des Konvertierungstools *SaltNPepper*[24] können Sie Korpora aus vielen gängigen Formaten in ein ANNIS-konformes Format überführen und dann mit ANNIS durchsuchen.

Das Web zum Korpus macht schließlich das browser-basierte *WebCorp*. Es handelt sich um ein kostenfreies Konkordanz-Tool, mit dem Sie linguistische Anfragen an das Internet stellen können. Die Suchergebnisse werden im KWIC-Format präsentiert[25]. In diesem Zusammenhang wollen wir Sie nochmals auf die Probleme mit dem World Wide Web hinweisen, die wir in Kapitel 3.1.3 diskutiert haben. Wenn Sie das Internet als Korpus nutzen wollen, sollten Sie diese reflektieren.

---

20 TüNDRA: `http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra`.
21 ANNIS: `http://annis-tools.org/`.
22 Demoversion: `http://korpling.german.hu-berlin.de/Annis/search.html`.
23 PostgreSQL: `http://www.postgresql.org/`.
24 ANNIS-Tools: `http://annis-tools.org/tools.html`.
25 WebCorp: `www.webcorp.org.uk/`.

5.1.3 Anfragen formulieren

Beim Formulieren von Suchanfragen kann es sein, dass Sie als Linguist ihre Perspektive auf die Daten ändern müssen. Eine erfolgreiche Korpusabfrage setzt voraus, dass die linguistischen Fragestellungen auf die Gegebenheiten des Korpus abgebildet wurden[26]. Man spricht hier von einer *Operationalisierung* der linguistischen Fragestellungen[27]. Ist das Korpus auf Wortebene annotiert, können Sie syntaktische Zusammenhänge z.B. in lineare Abfolgen von Wörtern und Wortarten übersetzen. Konzepte, die Ihnen hierbei zur Verfügung stehen, sind direkte und mittelbare *Präzedenz*, intervenierende Elemente, Optionalität, Alternation, Wiederholungen und Satzgrenzen (siehe auch den Exkurs zu regulären Ausdrücken, S. 92ff.). Im Folgenden spielen wir beispielhaft die Operationalisierung von zwei linguistischen Fragestellungen durch.

Die erste Fragestellung soll durch eine CQP-basierte Suche auf dem bereits im letzten Abschnitt erwähnten CQP-Demokorpus zu Bundestagsdebatten[28] umgesetzt werden. In Tabelle 8 sehen Sie zunächst eine linguistische Fragestellung zu Akkusativ-mit-Infinitiv-Verben (auch *AcI-Verben* von Latein ‚accusativus cum infinitivo'). Diese wurde operationalisiert und in eine Korpusanfrage übersetzt, welche sowohl in Worten ausbuchstabiert als auch in der Anfragesyntax (hier als CQP-Anfrage) formalisiert ist. Anschließend haben wir zu Illustrationszwecken noch einen Korpusbeleg abgedruckt.

Linguistische Fragestellung	Gibt es Gegenbeispiele zur These, dass im Deutschen AcI-Verben wie *sehen* oder *hören* im Perfekt immer in der Form des Ersatzinfinitivs auftreten (vgl. Meurers, 2005)? Ein Ersatzinfinitiv wäre *hat ... reden hören*, ein Gegenbeispiel wäre die Partizipkonstruktion *hat ... reden gehört*.
Übersetzung für die Anfrage	Suche nach einem Wort mit dem POS-Tag VVINF (*Vollverb im Infinitiv*) unmittelbar gefolgt von *gesehen* oder *gehört*.
CQP-Anfrage	`[pos = "VVINF"]("gesehen"|"gehrt")`
Korpusbeleg	„Ich habe gestern Frau Matthäus-Maier hier **reden gehört**, die leidenschaftlich dafür geworben hat, daß man die Schulden reduziert (...)"

Tabelle 8: Beispielanfrage zu AcI-Verben im CQP-Demokorpus

Steht einem ein phrasenstrukturell annotiertes Korpus zur Verfügung, kann man zusätzlich zur linearen Abfolge (siehe Präzedenz) über das Konzept der *Dominanz* auf hierarchische Beziehungen Bezug nehmen. Neben direkter und mittelbarer Dominanz kann man, falls das Korpus Kantentags enthält, auch gelabelte Dominanz miteinbeziehen.

---

[26] Vgl. auch Meurers und Müller (2008).

[27] Im Zusammenhang mit der quantitativen Auswertung von Korpora gehen wir in Kapitel 6.2.1 noch einmal ausführlich auf Operationalisierungen ein.

[28] CQP-Demokorpus: `http://corpora.linguistik.uni-erlangen.de/demos/CQP/cqpdemo.html`.

Das zweite Beispiel, das in Tabelle 9 dargestellt ist, illustriert eine Suchanfrage mit TIGERSearch auf der TüBa-D/Z[29]. Abbildung 13 zeigt zusätzlich die grafische Anfrageoption von TIGERSearch, bei der man Teilstrukturen ‚zeichnen' kann. Der hellgrau hinterlegte Bereich beschreibt die nicht-terminalen Knoten im Baum, der dunkelgrau hinterlegte die terminalen. Die Kanten zwischen den Knoten können mit funktionalen Tags markiert werden (hier z.B. *PRED* und *HD* für Prädikativ und funktionalen Kopf).

Linguistische Fragestellung	Gibt es Prädikativkonstruktionen, bei denen ein Genitiv als Prädikatsnomen fungiert?
Übersetzung für die Anfrage	Suche nach zwei nicht-terminalen Knoten (*#n1* und *#n2*), die in einem direkten Dominanzverhältnis zueinander stehen, wobei die verbindende Kante ein *PRED*-Tag trägt (= gelabelte Dominanz). Der untergeordnete Knoten (*#n2*) muss wiederum über ein *HD*-Tag (*HD* für *head, Kopf*) mit einem terminalen Knoten (*#n3*) verbunden sein, welcher als morphologische Markierung an erster Stelle ein *g* für Genitiv trägt.
Textuelle TIGERSearch-Anfrage	`#n1:[NT] >PRED #n2:[NT] &` `#n2 >HD #n3:[morph=/g.*/]`
Korpusbeleg	„Die Einsicht ist da, und die Opposition **ist** im Grunde **derselben Ansicht**."

Tabelle 9: Beispielanfrage *Genitivprädikativ* in der TüBa-D/Z

5.2 Eigenes Annotieren

In diesem Abschnitt wollen wir wie schon in Kapitel 3.4 wieder ein paar Tipps geben, für den Fall, dass Sie ein eigenes Korpus erstellen und selbst annotieren wollen. Dabei wollen wir Sie zunächst dazu anleiten, sich an bestehenden Standards zu orientieren. Dies ist nicht nur für eine potenzielle Weitergabe und Nachnutzung Ihrer Daten wichtig, sondern auch für den Einsatz von Tools, die Ihnen die Annotation oder die Auswertung erleichtern können. Darüber hinaus zeigt der Einsatz von Standards, dass Sie sich im Vorfeld über den State-of-the-Art informiert haben, was ein wichtiges Kriterium bei der Evaluierung von Projekten z.B. bei Promotionen oder beim Einwerben von Drittmittelprojekten darstellt. !

[29] Sie können die Anfrage alternativ auch in TüNDRA stellen, das wir im letzten Abschnitt beschrieben haben. Beachten Sie, dass die Anfrage nur einfache Fälle abdeckt. Für komplexere Nominalstrukturen müsste sie erweitert werden.

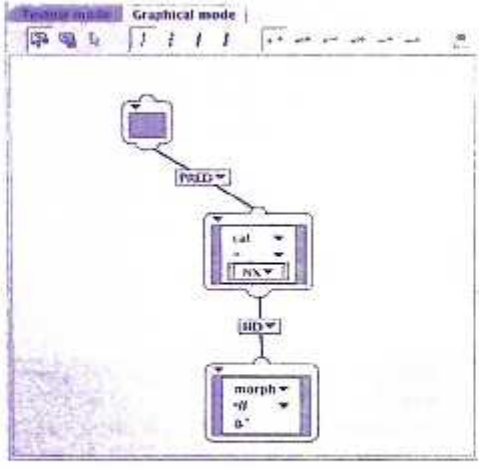


Abbildung 13: Grafische Anfrage in TIGERSearch: *Genitivprädikativ* in TüBa-D/Z

5.2.1 Standards

Als allgemeine Herangehensweise ans Annotieren wollen wir Sie auf die Annotationsmaximen nach Geoffrey Leech[30] verweisen (siehe Abb. 14), die wir im Anschluss kommentieren. Leech hatte beim Formulieren der Maximen größere Annotationsprojekte im Sinn. Wir denken aber, dass sie mit kleinen Einschränkungen auch für kleine Vorhaben einschlägig sind.

1. Annotation sollte so eingetragen sein, dass man den Ursprungstext wiederherstellen kann.
2. Es sollte möglich sein, die Annotation unabhängig vom Ursprungstext abzuspeichern und auszuwerten.
3. Die Annotation sollte dokumentiert werden, z.B. in der Form von Richtlinien. Die Dokumentation sollte dem späteren Nutzer zur Verfügung stehen.
   a) Die Annotatoren und die Annotationsumstände sollten bekannt sein.
   b) Die Qualität des Korpus sollte überprüft und dokumentiert werden. Die Benutzer sollten erfahren, wie konsistent die Annotation ist.
4. Das Annotationsschema sollte so weit wie möglich theorieneutral sein.
5. Kein Annotationsschema kann a priori als Standard gelten – Standards bilden sich durch Konsens der Nutzer heraus.

Abbildung 14: Annotationsmaximen nach Geoffrey Leech (1997)

[30] Vgl. Leech (1997).

Wie können Sie vorgehen, wenn Sie Leech folgen und die Annotation von den Rohdaten trennen wollen? Früher hat man die linguistischen Tags einfach an die zu beschreibenden Elemente angehängt, z.B. mit einem Unterstrich oder einem Schrägstrich wie in (10):

(10) ein/ART einfaches/ADJA Beispiel/NN

Zu jedem Wort wird hinter einem Strich die Wortklasse angegeben. Die Annotationstags sind damit eindeutig vom Primärtext getrennt und man kann bei Bedarf leicht nach Folgen von Wörtern und Wortarten suchen. Allerdings wird diese Art der Annotation schnell unübersichtlich, wenn man mehrere Eigenschaften gleichzeitig annotieren möchte. Deshalb ist man dazu übergegangen, für die Annotation spezielle Auszeichnungssprachen zu verwenden, wobei XML in den letzten Jahren eine zentrale Rolle spielt[31]. Die Auszeichnung mit XML besteht größtenteils aus Elementen, die ihrerseits aus öffnenden und schließenden Auszeichnern in spitzen Klammern gebildet werden z.B. <*w*> ... < /*w*>. Die Auszeichner werden wiederum als Tags bezeichnet[32]. Im folgenden Beispiel ist jedes Token von einem XML-Element (hier mit dem Namen w für Worttoken) umschlossen.

(11)
```
<w pos="ART">ein</w>
<w pos="ADJA">einfaches</w>
<w pos="NN">Beispiel</w>
```

Das schließende Tag wiederholt den Elementnamen mit einem vorangestellten Schrägstrich ('Slash'). Zusätzlich zu den Elementnamen können die öffnenden Tags Attribute beinhalten (hier z.B. *pos* mit dem Wert *ART*). Beispiel (13) zeigt noch eine Variante der Darstellung von XML-Elementen. Wenn ein Element wie in (12) „leer" bleibt, also kein Text zwischen dem öffnenden und schließenden Tag steht, kann man es auch verkürzt hinschreiben (13).

(12) `<s id="3"></s>`

(13) `<s id="3"/>`

Auch wenn man mittels des Markups von XML der Anforderung von Leech gerecht werden kann, Annotation und Ursprungstext sauber von einander getrennt zu verwalten[33], sind damit noch nicht alle Anforderungen an die Annotationskodierung erfüllen. Beispiel (14) stellt eine Erweiterung der Annotation von Beispiel (11) dar. Hier wurde eine zusätzliche nicht-terminale (*nt*) Annotationsschicht mit phrasaler Annotation (Phrasenkategorie *cat* = Nominalphrase *NP*) hinzugefügt.

(14)
```
<nt cat="NP">
<w pos="ART">ein</w>
```

---

[31] Die Abkürzung XML steht für *Extensible Markup Language*. Für eine Einführung in diesen Formalismus verweisen wir den interessierten Leser auf die Übersetzungen der XML-Standards `http://www.edition-w3c.de/` sowie auf das einführende Buch von Henning Lobin (2000).

[32] Sie kennen nun zwei Bedeutungen für das Lehnwort *Tag*: als Annotationslabel und als Teil eines XML-Elements.

[33] Siehe Punkt 1 und 2 in Abb. 14.

```
<w pos="ADJA">einfaches</w>
<w pos="NN">Beispiel</w>
</nt>
```

Stellen Sie sich nun vor, dass Sie die Annotationsschemata zweier Korpora vergleichen wollen, indem Sie einen Beispieltext sowohl nach dem TIGER-Korpus als nach der TüBa-DZ annotieren[34]. Bei einfachen Fällen könnte dies durch spezifische Attribute wie in (15) gelöst werden oder auch durch das Hinzufügen korpusspezifischer Elemente (z.B. nt_tiger, nt_tueba).

(15)
```
<nt cat_tiger="NP" cat_tueba="NX">
<w pos="ART">ein</w>
...
```

Diese Methode hat allerdings Grenzen. Wenn teilüberlappende Strukturen entstehen, ist es nicht möglich die konkurrierenden Spannen in einer gemeinsamen XML-Datei wie bisher gezeigt abzubilden. Ein Beispiel hierfür ist in Tab. 10 skizziert, in dem ein Substantiv durch einen Relativsatz modifiziert wird. Nach TIGER wird die Struktur als eine komplexe Nominalphrase (NP) annotiert, nach TüBa wird sie auf zwei topologische Felder verteilt (Mittelfeld MF und Nachfeld NF)[35]. Die beiden in der Tabelle dunkelgrau hinterlegten Spannen bilden eine teilüberlappende Struktur, in der keine der beiden Spannen vollständig von der anderen abgedeckt wird.

Token	Das	ist	für	mich	eine	Botschaft	,	die	den	...
TIGER-Phrasen			PP							
TüBa-Felder	VF	LK						NF		

Tabelle 10: Teilüberlappende Annotationseinheiten beim Vergleich des TIGER- und TüBa-Annotationsschemas von *Das ist für mich eine Botschaft, die den Deutschen jetzt gemäß ist.* (TIGER v2.1, Satz 3773)

Eine Lösung hierfür ist das sogenannte *XML-Standoff-Format*, bei dem die Annotationen getrennt vom Primärtext gespeichert werden und auf diesen nur verweisen. In (16) ist das einfache Beispiel von oben in einem Standoff-Format abgebildet. Anstelle verschachtelter XML-Elemente verweisen *Pointer*-Attribute (*idref*, *span*) auf die entsprechenden Stellen im Primärtext. Die TIGER und TüBa-Spannen sind dabei vollkommen unabhängig von einander und könnten entsprechend auch teilüberlappende Einheiten bilden.

(16)
```
<w id="w1">ein</w>
<w id="w2">einfaches</w>
<w id="w3">Beispiel</w>
<pos idref="w1">ART</pos>
```

[34] Vgl. die Gegenüberstellung in Tab. 5, S. 78.

[35] Für die Analysen s. Albert et al. (2003), S. 30f., und Telljohann et al. (2012), S. 99f.

```
<pos idref="w2">ADJA</pos>
<pos idref="w3">NN</pos>
<nr_tiger span="w1..w3">NP</nr_tiger>
<nr_tueba span="w1..w3">NX</nr_tueba>
```

Eine Speicherung im Standoff-Format ist immer auch dann sinnvoll, wenn man sich die Möglichkeit offen halten möchte, nachträglich weitere und ggf. konkurrierende Annotationsebenen zu einem Korpus hinzuzufügen. Normalerweise tippt man Standoff-Annotationen nicht selbst ein, sondern verwendet Annotationstools. Viele der Annotationstools, die wir Ihnen in Abschnitt 5.4 vorstellen werden, speichern die annotierten Daten in einem Standoff-Format ab.

Wie schon bei den Metadaten verweisen wir Sie wieder auf den *Corpus Encoding Standard* bzw. seine XML-Version *XCES*[36] als einen der klassischen Standards für die Kodierung von Annotation, d.h. für die Art und Weise, wie Annotationstags in einer Datei aufgeschrieben werden. PAULA und GRAF sind zwei generische XML-Formate, die von vielen Projekten für die Speicherung von Standoff-Annotation verwendet werden[37].

Inhaltliche Standards, d.h. Standards für die Namen (und Bedeutungen) der Annotationstags selber lassen sich nur schwer etablieren, da die Kategorien stark von den theoretischen Annahmen der Forscher abhängen. EAGLES[38] war ein früher Versuch, einen Konsens für verschiedene Annotationsebenen in Bezug auf die Analyse europäischer Sprachen zu erreichen. Um der Vielfalt an Annotationsschemata gerecht zu werden und dennoch eine Systematisierung zu erlauben, wurde die *ISOcat Registry*[39] eingerichtet, eine Webdatenbank, in der Annotationsschemata gesammelt werden und dadurch leichter zu vergleichen sind. Ein formaler Vergleich ist über explizite Bezüge in ISOcat oder über Metaschemata wie die *Ontologies of Linguistic Annotation* (OLiA)[40] möglich. Für die Annotation von Zeitausdrücken und -relationen hat sich das Annotationsschema *TIME-ML*[41] international durchgesetzt. *ISO-Space*[42] ist ein Versuch, die Annotation der Beziehungen zwischen Sprache und Raum, die z.B. durch Präpositionen ausgedrückt wird, zu standardisieren.

5.3 Entwicklung eines Annotationsschemas

Sollten Sie sich entschließen ein eigenes Annotationsschema zu entwickeln, wollen wir Ihnen methodische Hilfestellung an die Hand geben.

Die Entwicklung eines guten Annotationsschemas bzw. guter Annotationsrichtlinien verlangt, dass man die Kategorien sehr genau ausbuchstabiert. Der Hintergedanke hierbei ist, dass die Annotation im Idealfall nicht von Ihnen selbst, sondern von jemand

---

[36] XCES: `http://www.xces.org/`.
[37] Für PAULA s. Dipper (2005); Zeldes et al. (2013), für GRAF Ide und Suderman (2007).
[38] EAGLES-Empfehlungen: `http://www.ilc.cnr.it/EAGLES/browse.html`.
[39] ISOcat: `http://www.isocat.org/`.
[40] OLiA: `http://acoli.cs.uni-frankfurt.de/resources/olia/`.
[41] TIME-ML: `http://www.timeml.org/site/index.html`.
[42] ISO-Space: `https://sites.google.com/site/wikiisospace/`.

anderem durchgeführt wird. Dieses Vorgehen hilft sicherzustellen, dass die Annotationskategorien nachvollziehbar und objektiv definiert sind. In Abschnitt 4.3.1, S. 64f., hatten wir Ihnen bereits das Annotationsschema des Stuttgart-Tübingen Tagsets vorgestellt. Idealerweise besteht ein Schema bzw. die Richtlinien für die Annotatoren aus folgenden Bausteinen (nicht notwendigerweise in dieser Ordnung):

- Eine Liste aller Tagnamen (meist sprechende Kürzel) zusammen mit ihren Langnamen (Kategoriebezeichnungen),
- Definitionen der Kategorien,
- Prototypische Annotationsbeispiele für die Kategorien,
- Tests, die helfen zu entscheiden, ob eine Kategorie zutrifft,
- Problematische Beispiele mit Annotationen,
- Typische Verwechslungskategorien (d.h. konkurrierende Tags) mit Beispielen.

Tests für die Unterscheidung zwischen der Wortartenannotation ADJD und VVPP hatten wir Ihnen bereits in Abschnitt 4.3.1, S. 65f. vorgestellt[43]. Als weiteres Beispiel für einen Test zitieren wir aus dem TIGER-Annotationsschema. Es geht hier um die Entscheidung, ob eine Nominalphrase als Subjekt oder als Prädikativ annotiert werden soll:

> **Weitere Tests:**
>
> (i) Ersetze die Kopula durch eine Form von *machen*, das (vermeintliche) Subjekt durch eine Akkusativ-NP und das Prädikatsnomen durch eine zu-PP:
>
> (52) der Gärtner ist der Mörder
>
> a. sie haben den Gärtner zum Mörder gemacht
>
> b. ??????????sie haben den Mörder zum Gärtner gemacht (wenn schon, dann den Bock :-)
>
> (ii) noch besser: *gelten als, etw. darstellen, ...*
> → die *als-Phrase* ist das Prädikativ:
>
> (53) a. der Gärtner gilt als Mörder
>
> b. ≠ ??der Mörder gilt als Gärtner
>
> Auch wenn beide Möglichkeiten nicht besonders gut klingen, nimm die bessere! Falls immer noch unklar, kann das Label SP (subject or predicative) als ultima ratio vergeben werden.
>
> (Albrecht et al. 2003, S. 66, Fragezeichen im Original)

In der Praxis hat es sich bewährt, Annotationsentscheidungen systematisch in Entscheidungsbäumen zu organisieren vgl. das Beispiel auf S. 67. Gleichzeitig können auch die Annotationskategorien hierarchisch aufgebaut sein, so dass bei Unsicherheiten ggf. auf eine Oberkategorie zurückgegriffen werden kann – das muss aber nicht sein[44].

---

[43] Vgl. Schiller et al. (1999), S. 24.

[44] Ein Beispiel für ein hierarchisches Tagset und den Einsatz von Entscheidungsbäumen ist die Annotation von Präpositionenlesarten in einem Projekt von Tibor Kiss vgl. Kiss (2011); Müller (2013); Kiss et al. (2014).

Ein Annotationsschema sollte nicht am Reißbrett entworfen werden, sondern in Auseinandersetzung mit den zu annotierenden Daten. Es bietet sich an, hierfür eine Pilotphase bei der Korpusannotation einzuplanen. In Abb. 15 skizzieren wir den *Annotationszyklus*, in dem ausgehend von Literaturrecherche und erster Datensichtung (siehe *Daten &Theorie*) ein Schema entworfen wird, welches dann nach und nach – in einem iterativen Prozess – ausgebaut und verbessert wird. Die entscheidenden, sich wiederholenden Schritte sind (i) Analysieren von Daten beim Annotieren, (ii) Diskutieren bzw. Evaluieren der Annotationen, (iii) Interpretieren der Annotationsprobleme und entsprechende Erweiterung oder Revision des bisherigen Schemas.

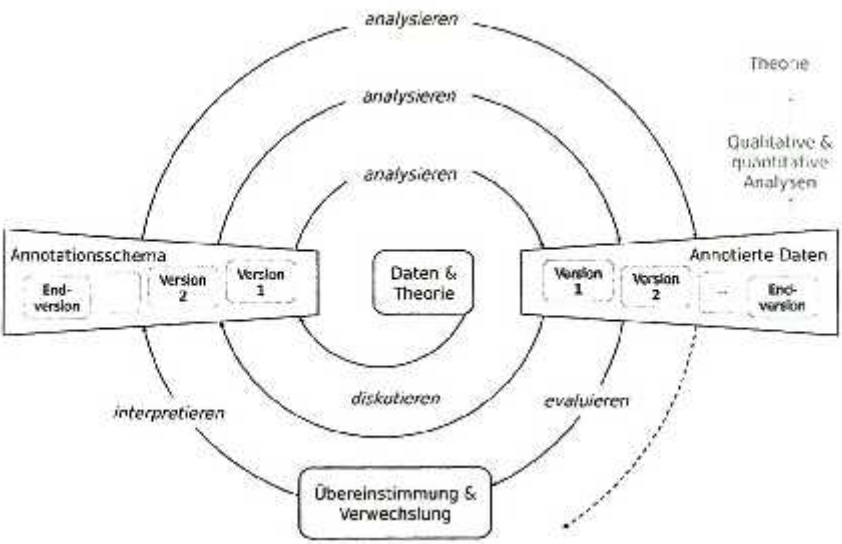


Abbildung 15: Der Annotationszyklus

Für die Evaluierung benötigt man Doppel- oder Mehrfachannotationen derselben Daten entweder durch mehrere Personen oder von einer Person zu verschiedenen Zeitpunkten. Differenzen weisen auf potenziell schwierige Phänomene bzw. problematische Kategorien hin, die im Annotationsschema (besser) abgedeckt werden müssen[45].

Verwechslungskategorien werden in einer *Verwechslungsmatrix* ermittelt. Tabelle 11 fasst beispielhaft die Wortartenannotationen von zwei Annotatoren (links) in einer Verwechslungsmatrix (rechts) zusammen.

[45] Die dritte Möglichkeit – dass die Annotatoren einfach schlecht sind – sollte man versuchen, ausschließen zu können.

Instanz	Annotator 1	Annotator 2
1	ADJD	ADJD
2	VVPP	ADJD
3	VVPP	VVPP
4	ADJA	ADJA
5	ADJD	VVPP
6	ADJD	ADJD
7	ADJD	ADJD
8	ADJA	ADJA

		Annotator 2		
		ADJA	ADJD	VVPP
Annotator 1	ADJA	2	0	0
	ADJD	0	3	1
	VVPP	0	1	1

Tabelle 11: Evaluierungsbeispiel mit Doppelannotation (links) und Verwechslungsmatrix (rechts)

Hier scheint Kategorie ADJA unkontrovers zu sein. ADJD und VVPP hingegen sind Verwechslungskandidaten. Ihre Definitionen und die Tests in den Annotationsrichtlinien sollten auf der Basis der Korpusbeispiele überprüft werden. Ggf. muss das Annotationsschema revidiert werden, indem gröbere, feinere oder einfach andersartige Kategorien definiert werden. Zusätzlich zu den beobachteten Übereinstimmungen und Differenzen gibt es Möglichkeiten, zufällige Übereinstimmung herauszurechnen[46]. Übereinstimmungsangaben, die im Anschluss an die Pilotphase beim eigentlichen Annotieren erhoben werden, können als Information über die Qualität der Annotation bzw. deren Konsistenz zusammen mit den Korpusdaten veröffentlicht werden[47].

In der Computerlinguistik beschrieb James Pustejovsky den zirkulären Entwicklungsverlauf von Annotation und davon abgeleiteten Analysetools als MATTER-Methode: ‚model', ‚annotate', ‚train', ‚test', ‚evaluate', ‚revise'[48]. Die Trainings- und Testschritte beziehen sich hierbei auf die Entwicklung von statistischen, computerlinguistischen Tools. Die restlichen vier, annotationsrelevanten Schritte werden zusammengefasst auch als MAMA-Methode bezeichnet.

Der Annotationszyklus ist ebenfalls angelehnt an den klassischen *hermeneutischen Zirkel* der textbezogenen Geisteswissenschaften: Man gelangt von einem Vorverständnis einer Theorie über Textverständnis zu einem verbesserten Theorieverständnis usw.[49]. Im Zuge der Digital Humanities wurde von Evelyn Gius, Janina Jacke, Jan Christoph Meister und Kollegen ein erweiterter hermeneutische Zirkel vorgeschlagen, der den Textdaten in der zweiten Runde des Zirkels Annotationen und später überprüfte Annotationen zur Seite stellt. Analog werden die Annahmen zur Theorie auf zweiter Ebene mit Richtlinien erweitert, in späteren Runden des Zirkels mit Richtlinien, Diskussionen und den vorhergehenden Analysen[50].

---

[46] Siehe z.B. Cohens *kappa* oder Krippendorffs *alpha* in Artstein und Poesio (2008). Siehe auch die Erläuterungen in Perkuhn et al. (2012).

[47] Vgl. Leechs Annotationskriterium Nr. 5.

[48] Siehe Pustejovsky und Stubbs (2012).

[49] Vgl. z.B. Gadamer (2010). Wir bedanken uns bei Janina Jacke, die uns darauf hinwies.

[50] Vgl. Bögel et al. (2015), S. 123.

5.4 Annotationstools

5.4.1 Manuelle Annotation

Um ein Korpus zu annotieren, benötigen Sie im Grunde keine besondere Software. In der Praxis haben sich aber spezialisierte Tools bewährt, die z.B. die Konsistenz der Annotation und die Fehlersuche unterstützen – und nicht zuletzt standardisierte Ausgabedateien erzeugen (Stichwort: XML). Wenn Sie ein kommerzielles Textverarbeitungsprogramm verwenden, empfiehlt es sich, die Daten in reinem Textformat abzuspeichern (.txt). Programmspezifisches Layout wie Kursivsetzung oder Fettdruck sind für die Annotation nicht geeignet, da diese Information verloren gehen könnte, wenn Sie das Format der Dateien in ein anderes Format konvertieren z.B., um die Dateien in ein anderes Programm einzulesen[51].

Spezielle Annotationstools vereinfachen die Annotation. Eine sehr allgemein Annotationsumgebung ist z.B. das Open-Source Programm *WordFreak*[52]. Ein anderes weit verbreiteres Tool ist das in Hamburg für die Transkription und Annotation von gesprochener Sprache entwickelte EXMARaLDA[53]. Es eignet sich für flache Annotationen jeglicher Art und erlaubt auch die Integration von multi-modalen Daten. In EXMARaLDA können Sie den *TreeTaggers* aufrufen und Ihre Texte automatisch mit STTS-Wortartentags annotieren. Für die Annotation von satzübergreifenden Phänomenen wie Koreferenzrelationen eignet sich das Heidelberger MMAX2[54] (gesprochen [maks] zwei, wie der Vorname). Manuelle Annotation von syntaktischen Konstituentenstrukturen wird von *Atomic*[55] unterstützt, das an der Friedrich-Schiller-Universität in Jena entwickelt wird. Das *RSTTool*[56] von Michael O'Donnell ist, wie der Name wahrscheinlich vermuten lässt, eine Annotationsumgebung für die manuelle Analyse von Diskursrelationen mit der *Rhetorical Structure Theory* (RST). Es erlaubt aber auch, dass man als Nutzer eigene Relationstypen definiert.

Zuletzt wollen wir Sie noch auf Tools verweisen, die neben einer lokalen Installation auch die Nutzung über einen Server erlauben und damit für *kollaboratives* Online-Annotieren einsetzbar sind. *Arborator*[57] wurde von Kim Gerdes an der Sorbonne Novelle in Paris für die Annotation von syntaktischen Dependenzstrukturen entwickelt. *WebAnno*[58] und *brat*[59] eignen sich für jegliche wortbasierte oder relationale Annotation (d.h. z.B. syntaktische Dependenzen oder Koreferenzrelationen), allerdings nicht für die Annotation von Konstituenten. Die Annotationsplattform *GATE*[60] wird bereits seit vielen Jahren an der Universität Shetfield in England entwickelt. Der Quellcode ist wie bei vielen der hier angeführten Tools open-source, so dass sich weltweit Programmierer an

---

51 Vgl. Bird und Simons (2003).
52 WordFreak: `wordfreak.sourceforge.net`.
53 EXMARaLDA: `www.exmaralda.org`.
54 MMAX2: `http://mmax2.net`.
55 Atomic: `http://linktype.iaa.uni-jena.de/atomic/`.
56 RSTTool: `\http://www.wagsoft.com/RSTTool/`.
57 Arborator: `http://arborator.ilpga.fr/`.
58 WebAnno: `https://www.ukp.tu-darmstadt.de/software/webanno/`.
59 brat: `http://brat.nlplab.org/`.
60 GATE: `https://gate.ac.uk/overview.html`.

Weiterentwicklungen beteiligen können. GATE ist sehr gut dokumentiert und bietet u.a. die Option, automatische Tagger und andere Tools zu integrieren.

Wie oben bereits angesprochen, werden nicht nur in der Linguistik Annotationen erstellt. Im Zuge der Digital Humanities etabliert sich das Annotieren in vielen geisteswissenschaftlichen Disziplinen. Wir wollen Sie hier auf ein Tool aus den Literaturwissenschaften aufmerksam machen, das Jan Christoph Meister aus Hamburg für die narratologische Analyse von Texten entwickeln ließ. *CATMA*[61] lässt sich intuitiv bedienen und ist durchaus auch für linguistische Annotationsvorhaben geeignet, wenn Sie nur einzelne Wörter oder Sequenzen mit Annotationstags auszeichnen wollen.

5.4.2 Automatische Annotation

Neben dem eben erwähnten GATE wollen wir hier nur auf eine weitere Online-Plattform für die automatische Annotation von Texten verweisen. Im deutschlandweiten Verbundprojekt CLARIN entstand die Online-Plattform WebLicht[62], die für akademische Nutzer nach Anmeldung kostenlos ist (siehe im Aufgabenteil am Ende des Kapitels). WebLicht versammelt eine ganze Reihe von linguistischen Annotationstools u.a.

- den TreeTagger für die Wortarten-Annotation mit STTS;
- den Berkeley-Parser, der syntaktische Konstituentenstrukturen einschließlich der Angabe von topologischen Feldern erzeugt wie in der TüBa-D/Z-Baumbank;
- dem BitPar-Parser[63] (in WebLicht als *Stuttgart Constituent Parser* bezeichnet), der eine hybride Konstituentenstruktur á la TIGER-Korpus annotiert, wobei die Phrasentags zusammen mit funktionalen Tags als komplexe Knotennamen erscheinen z.B. *NP-SB* für die Subjektsnominalphrase;
- den MATE-Parser[64] (in WebLicht *Stuttgart Dependency Parser* genannt), der syntaktische Dependenzannotationen ausgibt, die von den funktionalen Kanten des TIGER-Korpus abgeleitet sind.

! Ein wichtiger Grundsatz bei der Nutzung von automatischen Annotationstools ist, dass immer nur das ausgegeben werden kann, was durch Regeln oder sogenannte Trainingsdaten in das Tool hineingegeben wurde. Ein Wortartentagger, der nur auf Zeitungstext, aber nicht auf Chat- oder anderen internetbasierten Daten entwickelt wurde, wird keine passenden Annotationen für Smileys oder die Abkürzung *lol* (laugh out loud[e]) vorschlagen. Die Wortarten-Ausgaben des TreeTaggers unterscheiden sich je nach Sprachvariante des Tools[65]. Der deutsche TreeTagger annotiert Artikel mit dem Tag *ART* (vgl. STTS-Tagset), der englische TreeTagger mit dem Tag *DT* nach dem PennTreebank-Tagset und der französische TreeTagger gibt *DET:ART* aus. Die Tags sind jeweils durch das Korpus bestimmt, von dem der TreeTagger die Auftretenshäufigkeiten und Wort/Tag-Sequenzen gelernt hat.

---

61 CATMA: `http://www.catma.de/`.

62 WebLicht: `weblicht.sfs.uni-tuebingen.de/weblichtwiki/`.

63 BitPar-Parser: `http://www.cis.uni-muenchen.de/~schmid/tools/BitPar/`.

64 MATE-Parser: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.html`.

65 TreeTagger: `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

5.5 Weiterführende Literatur

Einen tieferen Einblick in Probleme und Lösungen der konkreten Korpuserstellung erhalten sie bei Sasaki und Witt (2004). Die Autoren betonen Aspekte der Texttechnologie, d.h. Fragen der konkreten Datenspeicherung und in welchem Format Annotationen in ein Korpus integriert werden können.

Zur automatischen Vorverarbeitung und Annotation mit computerlinguistischen Werkzeugen möchten wir Ihnen ebenfalls noch Leseempfehlungen geben. In der Einführung *Computerlinguistik und Sprachtechnologie* (Carstensen et al., 2010) sind insbesondere Kapitel 3.4 und 3.5 relevant. Dort erfahren Sie, wie Tagger und Parser arbeiten. Kapitel 4.1 und 4.2 handeln von Korpora sowie Baumbanken und sind ebenfalls zu empfehlen, auch wenn Ihnen manches schon bekannt vorkommen wird. Zinsmeister (2015) diskutiert Chancen und Grenzen automatischer Annotation.

Zum Abschluss möchten wir noch auf die Studienbibliografie zur Computerlinguistik hinweisen (Cramer und Schulte im Walde, 2006). Dort finden Sie diverse Informationen zu Korpora, Annotationswerkzeugen und anderen computerlinguistischen Ressourcen (`www.coli.uni-saarland.de/projects/stud-bib/`).

5.6 Aufgaben

1. Suche auf dem DWDS-Kernkorpus nach Formen von *einen*[66].
   a. Öffnen Sie die Ressourcenseite des DWDS in einem Browser und machen Sie sich mit dem DWDS-Kernkorpus vertraut:
   `http://www.dwds.de/ressourcen/korpora/`
   Lesen Sie auch die detaillierte Darstellung auf der weiterführenden Seite (`http://www.dwds.de/ressourcen/kernkorpus/`): Wie viele Token umfasst das Kernkorpus 20? Welche Register beinhaltet es?
   b. Öffnen Sie dann die Start-Webseite des DWDS: `http://www.dwds.de/`
   c. Tippen Sie *einen* in das Suchfeld und drücken Sie die Enter-Taste: `einen` (Enter)
   Stellen Sie sicher, dass Sie die Darstellung ‚DWDS Referenzkorpora' aktiviert haben (siehe Drop-Down-Button rechts neben dem Lupe-Icon).
   d. Wie viele Verbvorkommnisse werden Ihnen im Panel des Kernkorpus 20 angezeigt, ohne dass Sie weiter scrollen? Wie viele im Panel des Deutschen Textarchivs?
   e. Schränken Sie die Suchanfrage zunächst auf die Wortform *einen* ein[67]: `@einen` (Enter). Sehen Sie nun Verbformen?
   f. Wie sieht es mit der Wortform *geeint* aus? Suche: `@geeint` (Enter).
   g. Machen Sie sich die Wortartenannotation zu nutze, um verschiedene Flexionsformen des Verbs zu finden: `einen with $p=VVFIN` (Enter)
   Das Ergebnis ist leider ziemlich ernüchternd. Bedenken Sie bei der Trefferdarstellung, dass das Korpus rein automatisch annotiert wurde und die Artikelsart vom Tagger stark präferiert wird.

---

[66] An dieser Stelle vielen Dank an Jessica Sohl für ihre konstruktiven Hinweise zu den Aufgaben.

[67] Lassen Sie sich nicht davon irritieren, dass das @-Zeichen hier etwas „verdruckt" aussieht.

Abbildung 16: DWDS – Startseite für die Suche auf den DWDS-Korpora

h. Schränken Sie die Suche durch den Kontext ein, indem Sie verlangen, dass das Suchwort vor einem Modalverb steht. Beachten Sie, dass Sie bei einer Suche nach einer Wortsequenz (hier zwei Wörter) immer doppelte Anführungsstriche setzen müssen: `"einen with $p=VV* $p=VM*"` (Enter)

i. Erweitern Sie zuletzt die Suche noch auf die Alternative, dass das Verb vor einem Satzendezeichen steht:
`"einen with $p=VV* $p=VM*" || "einen with $p=VV* $p=\$."` (Enter)

2. Suchen Sie im DWDS-Kernkorpus nach Gegenbeispielen zur These, dass im Deutschen Acl-Verben wie *sehen* oder *hören* im Perfekt immer in der Form des Ersatzinfinitivs auftreten (vgl. Abschnitt 5.1.3).

a. Öffnen Sie die Start-Webseite des DWDS `http://www.dwds.de/`, tippen Sie die Suchanfrage ein und drücken Sie die Enter-Taste.
`"$p =VVINF @gesehen" || "$p =VVINF @gehört"`
In Worten: Suche nach einem Satz, der ein Wort mit dem POS-Tag VVINF (*Vollverb im Infinitiv*) enthält, auf das unmittelbar die Wortform *gesehen* folgt oder (alternativ) nach einem Satz, der ein Wort mit dem POS-Tag VVINF (*Vollverb im Infinitiv*) enthält, auf das unmittelbar die Wortform *gehört* folgt. Achten Sie auf die korrekte Setzung der Anführungsstriche. Zur Erklärung der Syntax siehe wie oben `http://www.dwds.de/hilfe/suche/`.

Abbildung 17: DWDS – Menüauswahl für die Darstellung des Wortverlaufs

b. Sichten Sie die Suchergebnisse, indem Sie auf einzelne Treffer klicken und sich den Kontext und die Quellangaben anzeigen lassen. Handelt es sich um verlässliche Belege?

c. Klicken Sie rechts unten am Panelrahmen auf den Link *Optionen* und aktivieren Sie im Menüfenster die Option *Textsorte*. Gehen Sie zurück auf die Ergebnisliste. Wenn Sie mit der Maus auf das Textsortenkürzel gehen, wird Ihnen die Textsorte als Mouse-over-Effekt angezeigt. Überwiegt eine der Textsorten?

d. Klicken Sie rechts oben im Panelrahmen auf das Download-Icon, um die Ergebnisliste als Textdatei auf den eigenen Rechner zu speichern. Wenn Sie dort Zugriff auf die Textsortenklassifikation haben wollen, müssen Sie dies vor dem Download im Menü aktivieren. Die Textdatei können Sie sich für weiterführende Untersuchungen z.B. in ein Tabellenprogramm wie MS Excel importieren.

e. Abschließend wollen wir Ihnen noch ein weiteres Feature der DWDS-Seite vorstellen. Klicken Sie oben auf der Gesamtseite auf den Button *+Ressourcen* und wählen Sie zuerst *Statistiken* und dort *Wortverlauf (Basis DWDS-Kernkorpus)* aus. Sie erhalten dann ein zusätzliches Panel mit einer grafischen Darstellung der Treffer verteilt über die Dekaden des 20sten Jahrhunderts. Dieses Panel bietet wieder eine Reihe von Optionen an, die Sie durchtesten können. Unter anderem können Sie das Panel über den Button *Chart Context Menu* in verschiedenen Bildformaten herunterladen. Der Button versteckt sich in der rechten oberen Ecke des Panels und ist mit drei kleinen waagerechten Balken visualisiert. Bitte beachten Sie, dass sich die Datengrundlage für den Wortverlauf nur aus einer Basisversion des eigentlichen Kernkorpus 20 speist, so dass Sie hier teilweise

geringere Frequenzdaten sehen. Eine gewisse Vorstellung der diachronen Entwicklung erhalten Sie allemal.

3. Suchen Sie in der Dependenzversion der TüBa-D/Z, die über TüNDRA online zur Verfügung steht nach Beispielen für Prädikativkonstruktionen im Genitiv (vgl. Abschnitt 5.1.3).

Abbildung 18: TüNDRA – Auswahl der TüBa-D/Z Dependency

a. Öffnen Sie die die TüNDRA-Seite in einem Browser:
`http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra`.
b. Klicken Sie auf den Link `https://weblicht.sfs.uni-tuebingen.de/Tundra` und melden Sie sich im Clarin EU Service Provider über Ihre Universität an. Sollte Ihre Universität nicht aufgelistet sein (was z.B. zur Zeit der Überarbeitung des vorliegenden Buches für die Universität Hamburg galt) können Sie sich bei `https://user.clarin.eu/user/register` ein eigenes Konto einrichten lassen, mit dem Sie Zugriff auf TüNDRA über den *clarin.eu website account* erhalten. Es ist wichtig, dass Sie hier eine akademische E-Mail-Adresse angeben, da Ihre Anfrage andernfalls als Spam aussortiert wird.
c. Auf der TüNDRA-Startseite klicken Sie ganz oben links auf *Load Treebank* und wählen dann aus der Liste die TüBa-D/Z v9 Dependency (Experimental) aus. Achten Sie auf den korrekten Namen des Korpus!
d. Tippen Sie in das *Search*-Feld ganz oben ein:
`#n1 >PRED #n2:[morph=/g.*/]`
In Worten: Suche nach zwei (Wort-)Knoten #n1 und #n2, die mit einer PRED-Relation (Prädikativ) verbunden sind, wobei das abhängige Dependens mit dem Kasus Genitiv annotiert ist.

c. Sie sollten 73 Treffer erhalten. Machen Sie sich ein Bild von den Ergebnissen, indem Sie mit den Pfeiltasten oben in der Mitte durch die Treffer browsen. In der Icon-Leiste links werden Ihnen verschiedene Darstellungsoptionen angeboten. Klicken Sie sich einmal durch die Varianten. Welche Darstellungen helfen Ihnen, die Ergebnisse besser zu verstehen? Handelt es sich tatsächlich um Prädikativkonstruktionen mit Kopula oder würden Sie die Belege anders klassifizieren?

4. Artikelsetzung ist ein schwieriges Phänomen für Fremdsprachenlerner. Suchen Sie nach falsch verwendeten Artikeln im Lernerkorpus Falko. Gehen Sie dafür auf die ANNIS-Demo-Seite: `korpling.german.hu-berlin.de/annis3`

Abbildung 19: ANNIS – Suche auf dem Falko-Korpus

a. Wählen Sie in der *Corpus List* links unten das Korpus falkoEssayL2v2.4 aus. Es beinhaltet 248 Texte von Fremdsprachenlernern des Deutschen.

b. Tippen Sie in das Anfragefenster rechts oben eine Suche nach Artikeln ein, die auf der Ebene der Zielhypothese 1 (ZH1) gelöscht wurden:
`ZH1Diff="DEL" & #1 _=_ pos="ART"`
In Worten: Suche auf der Annotationsebene ZH1Diff nach Annotationen vom Typ DEL ('delete') und gebe genau solche Treffer (#1) aus, für die gilt, dass sie auf der Ebene der Wortartenannotation des Originals (pos) mit ART getaggt sind. Sie müssten 61 Treffer erhalten.

c. Browsen Sie durch die Ergebnisliste. Erweitern Sie bei Bedarf die Kontextsicht auf mehr als 5 Token (rechts oder links). Überprüfen Sie auch die „ZH1(grid)-Darstellung". Betrifft die Tilgung nur den Artikel oder handelt es sich um größere Korrekturen?

6 Den Wald hinter den Bäumen sehen — Quantitative Auswertung von Korpusdaten

Am Ende dieses Kapitels haben Sie anhand einer Beispielstudie gesehen, wie eine linguistische Fragestellung umformuliert werden kann, um sie „korpustauglich" zu machen. Sie haben ebenso gesehen, wie aus einer linguistischen Fragestellung eine quantitativ überprüfbare Hypothese abgeleitet wurde. Sie verstehen, wie man die Ergebnisse einer Korpusabfrage tabellarisch präsentiert, so dass Sie daraus verschiedene quantitative Zusammenfassungen und grafische Darstellungen erzeugen können. Sie haben anhand von zwei Beispielstudien Kennwerte für zentrale Tendenzen kennengelernt. Sie werden erkannt haben, wie wichtig die Visualisierung ist, um quantitative Daten zu verstehen. Außerdem haben Sie Hinweise auf Internetressourcen und weiterführende Literatur zur statistischen Auswertung von quantitativen Daten erhalten, da wir Ihnen im Rahmen dieses Buches nur einen Ausblick auf die eigentlichen Berechnungen geben können.

6.1 Korpuslinguistik und Statistik

Wie wir in den bisherigen Kapiteln gezeigt haben, befasst sich Korpuslinguistik mit Korpora, d.h. mit Primärdaten, Annotationen und Metadaten. Es geht dabei darum, wie man Korpora inhaltlich konzipiert, konkret erstellt und schlussendlich in Dateien in bestimmten Datenformaten repräsentiert und über Anfrageschnittstellen zugänglich macht. Ein anderer Bereich von Korpuslinguistik handelt davon, wie Korpusdaten für die linguistische Argumentation genutzt werden können. Hierzu gehören methodische Aspekte z.B., wie man eine Abfrage an ein Korpus stellt, um relevante Beispieldaten zu finden oder Überlegungen dazu, wie man die Korpusbeispiele angemessen interpretiert. Da Korpora im Idealfall Zugriff auf große Datenmengen bieten, befasst sich ein weiterer Teilbereich der Korpuslinguistik mit quantitativen Auswertungen. Hier kommen unweigerlich statistische Methoden ins Spiel. Dieses Kapitel soll anhand von einfachen Beispielen in die Denkweise und die ersten Schritte einer quantitativen Auswertung einführen[1]. Im Rahmen dieses Buches können wir keine Einführung in die Statistik geben, wir können die Leser nur anregen, sich tiefer mit diesem Thema zu befassen. Erfreulicherweise wurden in den letzten Jahren mehrere Statistikeinführungen veröffent-

[1] Herzlichen Dank an Melanie Andresen und Fabian Barteld für Korrekturen und konstruktive Vorschläge zu diesem Kapitel.

licht, die sich speziell an Linguisten wenden. Wir werden am Ende des Kapitels im Abschnitt zur weiterführenden Literatur kurz darauf eingehen.

6.2 Operationalisierung und Hypothesen

6.2.1 Operationalisierungen von Fragestellungen

Möchte man quantitativ arbeiten, muss man Dinge auszählen. In Bezug auf quantitative Korpusstudien setzt dies voraus, dass die linguistischen Einheiten, die man zählen möchte, im Korpus wiederauffindbar sind. Linguistische Fragestellungen müssen dafür auf die Gegebenheiten des Korpus angepasst werden. Sie müssen *operationalisiert* werden. Das klingt zunächst etwas sperrig, bedeutet aber nur, dass man die Konzepte einer linguistischen Fragestellung in Bezug auf ihre Auffindbarkeit im Korpus überprüft und, wenn nötig, auf beobachtbare Einheiten abbildet.

Die einfache linguistische Aufgabe „Zähle alle Adjektive im Korpus" setzt zum Beispiel voraus, dass der Begriff *Adjektiv* operationalisiert wird. Soll die Suche in einem nicht-annotierten Korpus stattfinden, benötigt man Kriterien, die eindeutig festlegen, was ein Adjektiv ist und was nicht. Man kann sich hierbei zum Beispiel auf eine bestimmte Grammatik oder auf vorhandene Annotationsrichtlinien berufen. Wenn man die Suchanfrage nicht operationalisiert und die Auffindungskriterien damit im Unklaren belässt, kann man schlussendlich auch keine aussagekräftigen quantitativen Ergebnisse ableiten.

Legt man der Auszählung ein annotiertes Korpus zugrunde, muss man überprüfen, in wie weit die Annotationskategorien das gesuchte Phänomen abbilden. Ist das Korpus zum Beispiel mit den Wortartentags des Stuttgart-Tübingen Tagsets (STTS) getaggt, stehen zwei Adjektivtags zur Verfügung und eine auf der Hand liegende Operationalisierung von *Adjektiv* wäre „Token mit den STTS-Annotationen ADJA oder ADJD", d.h. ein Token, das mit einem der beiden Tags für attributiv bzw. prädikativ/adverbial verwendete Adjektive getaggt ist. Bei einer entsprechenden Suchanfrage findet man dann alle Vorkommnisse, die gemäß STTS als ADJA oder ADJD annotiert wurden wie die Belege in (1) und (2).

(1) Wir haben gültige$_{\text{ADJA}}$ Pässe.

(2) Die bisherige$_{\text{ADJA}}$ Währung bleibt parallel$_{\text{ADJD}}$ als Zahlungsmittel gültig$_{\text{ADJD}}$.

Beispiel (3) zeigt, dass diese Operationalisierung eventuell noch nicht alle Vorkommnisse erfasst, die man gerne abdecken möchte. Kardinalzahlen wie *zwei* werden gemäß STTS als CARD ausgezeichnet – anders als Ordinalzahlen wie *zweite*, die das Tag ADJA erhalten[2].

(3) Mein Vater hatte zwei$_{\text{CARD}}$ Gesichter, und das zweite$_{\text{ADJA}}$ Gesicht war meistens verborgen.

Möchte man auch Vorkommnisse wie *zwei* in die Untersuchung einbeziehen, musste die Operationalisierung von *Adjektiv* lauten „Token mit den STTS-Annotationen AD-

[2] Vgl. Schiller et al. (1999), S. 18–28.

JA, ADJD oder CARD". Allerdings würde diese Anfrage auch Kardinalzahlen in nichtattributiver Verwendung auffinden wie z.B. Jahreszahlen. Man müsste die Operationalisierung noch einmal überarbeiten oder die Suchergebnisse nachträglich filtern.

Sollte das linguistische Phänomen, das man untersuchen möchte, keine unmittelbare Entsprechung in den Annotationstags haben, kann man sich ihm indirekt über leicht zu identifizierende Oberflächenmerkmale und relevante Annotationsmuster annähern. Da ein Fremdwort für *annähern* das Wort *approximieren* ist, bezeichnet man Stellvertretereinheiten für das eigentliche Phänomen als *Proxys*. Die Operationalisierung nimmt dann auf die entsprechenden Proxys Bezug.

Um die Verwendung von Proxys zu veranschaulichen, betrachten wir die quantitative Untersuchung des Informationsstatus von referenziellen Ausdrücken, vgl. Abschnitt 4.3.4. Zur Erinnerung, der Informationsstatus gibt grob gesagt an, ob ein Ausdruck auf einen Referenten verweist, der im Diskurs bereits eingeführt wurde (diskurs-alt und damit auch hörer/leser-alt) oder ob es sich um einen bisher unbekannten Referenten handelt (hörer/leser-neu und damit auch diskurs-neu). Die gängingen Analyseschemata sehen auch Mischformen und andere Unterklassen vor z.B. diskurs-neue Ausdrücke, die aber hörer/leser-alt sind, da das Wissen um ihre Existenz zum Allgemeinwissen gehört[3].

Tabelle 12 illustriert die Operationalisierung von Kategorien des Informationsstatus in Form von einfachen Proxys[4], die in einem syntaktisch annotierten Korpus per Suchanfragen auffindbar wären.

Kategorie	Kommentar	Proxy
Hörer-alt aber diskurs-neu	Allgemein bekannte Referenten	Nicht-vorerwähnte Eigennamen: z.B. Erstnennung der Stadt *Hamburg* im Text
Diskurs-alt	Referent ist im Text bereits eingeführt	Pronomen, vorerwähnte Eigennamen: z.B. *sie*; Zweitnennung von *Hamburg*
Hörer-neu	Werden als „brandneu" bezeichnet	Eigennamen, die von Relativsatz oder Apposition begleitet werden: z.B. *Peter Jackson, der Regisseur von „Herr der Ringe"*

Tabelle 12: Operationalisierung von Kategorien des Informationsstatus durch Proxys

Ähnlich, wie wenn man Annotationskategorien zur Operationaliserung eines linguistischen Phänomens verwendet, muss man bei der Operationalisierung durch Proxys immer hinterfragen, in wie weit man dem eigentlichen Untersuchungsphänomen gerecht wird. Referiert wirklich jedes Pronomen auf einen im Text vorerwähnten Referenten?

[3] Für Schemata zur Annotation von Informationsstatus siehe z.B. Prince (1981) oder Baumann und Riester (2012).

[4] Vgl. Strube und Hahn (1999).

Ganz sicher nicht. Das zeigen nicht-referierende Beispiele des Personalpronomens *es* in (4)–(6).

(4) *Formales Subjekt*
Es gibt zwei neue Maschinen auf dem Markt.

(5) *Korrelat des Objektsatzes*
Ich finde es gut, dass das öffentlich diskutiert wird.

(6) *Vorfeld-'es'*
Es nahmen drei Vertreter des Senats teil.

Man kann anhand einer Stichprobenauswertung abschätzen, in wie weit sich die Verteilung der Proxys im Text der tatsächlichen Verteilung des zu untersuchenden Phänomens annähert.

Im Folgenden betrachten wir wieder die Operationalisierung einer Aussage wie am Anfang dieses Unterkapitels. Anstelle der einfachen Aufforderung *Zähle alle Adjektive im Korpus*, bei dem wir das Konzept *Adjektiv* operationalisiert haben, handelt es sich nun um die Hypothese, *dass Fremdsprachenlerner des Deutschen (L2-Sprecher) Probleme mit der Sprache haben*. Neben rein substantivischen Termini wie *L2-Sprecher* (bzw. vorher *Adjektiv*), müssen auch die benannten Eigenschaften wie *Probleme haben* operationalisiert werden – was manchmal gar nicht so offensichtlich ist. Als Beispiel hierfür betrachten wir eine Untersuchung zur Informationsstruktur in Texten von Fremdsprachenlernern des Deutschen. Die Informationsstruktur beschreibt, wie alte und neue Information im Satz präsentiert wird. Der oben erwähnte Informationsstatus von Nominalphrasen ist ein Teilaspekt der Informationsstruktur. Wir legen die Hypothese zugrunde, dass L2-Sprecher Probleme mit der Informationsstruktur haben und zwar auch solche Lerner, die in Bezug auf Wortschatz und Kerngrammatik bereits sehr fortgeschritten sind. *Probleme haben* kann operationalisiert werden als „unterscheiden sich von L1-Sprechern des Deutschen". Dann wäre eine zu untersuchende These hierzu „L2-Sprecher verwenden eine andere Informationsstruktur als L1-Sprecher des Deutschen". Der Terminus *Informationsstruktur* ist abstrakt und das zugrundeliegende Phänomen sehr umfassend. Eine notwendige Herangehensweise ist, zunächst den Untersuchungsbereich klar einzugrenzen. In Bezug auf Informationsstruktur wäre es z.B. interessant, den Satzanfang bzw. linguistisch präziser, das Vorfeld ins Zentrum der Untersuchung zu stellen. Die Operationalisierung umfasst dann die Einschränkung auf einen bestimmten topologischen Bereich im Satz. Dieser Operationalisierungsschritt legt die Untersuchungsinstanzen fest: Die Studie wird Eigenschaften für einzelne Vorfeldinstanzen erheben.

Die zu überprüfende These würde dann lauten: „L2-Sprecher verwenden das Vorfeld anders als L1-Sprecher des Deutschen". Was hier noch fehlt – und das ist ganz entscheidend – ist die Operationalisierung von *anders*. Sie bestimmt schlussendlich, welche Merkmale wir in der Untersuchung berücksichtigen und auszählen werden. Vorfelder können *anders* besetzt werden in Bezug auf die syntaktische Funktion (Subjekt, Objekt, ...), die syntaktische Kategorie (Nominalphrase, Adverbphrase, ...), das „Gewicht" z.B. als Wortanzahl operationalisiert[5] oder das Gewichtsverhältnis zwischen Vorfeld und

[5] Ein schönes Beispiel dafür, welche Konsequenzen verschiedene Operationalisierungen von „Gewicht" haben, finden Sie in Gries (2008), S. 24.

dem restlichen Satz und vieles mehr. Die Operationalisierung von *anders* resultiert somit in der Festlegung eines oder mehrerer Merkmale (hier z.B. der syntaktischen Funktion), dessen konkrete Ausprägung für jede Untersuchungsinstanz dokumentiert wird (z.B. Subjekt oder Objekt). Statistisch ausgedrückt stellt das Merkmal eine *Variable* dar, weil sie unterschiedliche Ausprägungen annehmen kann. Auf das Konzept der Variable gehen wir in Abschnitt 6.3 noch einmal genauer ein.

6.2.2 Hypothesen bilden

Im letzten Abschnitt wurde gezeigt, wie man bei der Operationalisierung eine linguistische Fragestellung auf wiederauffindbare und damit zählbare Einheiten und Eigenschaften abbildet. Damit eng verbunden ist die Bildung von Hypothesen. Eine *Hypothese* ist die Umformulierung einer Fragestellung in eine Aussage, die durch eine empirische Untersuchung überprüft, d.h. im Zweifelsfall widerlegt werden kann. Im Fall von korpuslinguistischen Studien geschieht dies unter Berücksichtigung der durch die Operationalisierung festgelegten, zählbaren Einheiten und Eigenschaften. Zur Illustration gehen wir zurück zur Untersuchung der Informationsstruktur in Texten von Fremdsprachenlernern des Deutschen (L2-Sprechern).

- *Fragestellung:*
  Beherrschen fortgeschrittene L2-Sprecher die Informationsstruktur des Deutschen?
- *Operationalisierung:*
  L2-Sprecher verwenden andere syntaktische Funktionen im Vorfeld als L1-Sprecher des Deutschen.
- *Hypothese:*
  Die Häufigkeit der einzelnen Ausprägungen der Variable *Funktion* unterscheidet sich bei L2- und L1-Sprechern des Deutschen.

Hypothesen sind Behauptungen über die Ausprägungen einer Variable oder die Beziehung(en) zwischen zwei oder mehr Variablen in einem bestimmten Kontext. Sie enthalten Formulierungen wie „die gleiche Häufigkeit wie", „sind gleich", „unterscheiden sich", „je mehr / größer / ..., desto mehr / weniger / größer / ...".

In der Statistik spricht man von der sogenannten *Nullhypothese* H_0, der eine *Alternativhypothese* H_1 gegenübergestellt wird. In den meisten Fällen versucht man, die Nullhypothese zu widerlegen. Wichtig hierbei ist, dass die beiden Hypothesen so formuliert sind, dass man, wenn man die Nullhypothese verwirft, automatisch folgern kann, dass stattdessen die Alternativhypothese gilt. Dies hat zur Folge, dass die beiden Hypothesen nebeneinander gestellt oftmals etwas redundant wirken. Eine einfache Nullhypothese im Rahmen des Lernerkorpusbeispiels wäre die folgende, die anders als oben nur eine bestimmten Ausprägung der Variable *syntaktische Funktion* thematisiert:

- *Nullhypothese* H_0: In L2-Texten ist die Häufigkeit der Funktion Subjekt im Vorfeld gleich groß wie in L1-Texten.
- *Alternativhypothese* H_1: In L2-Texten ist die Häufigkeit der Funktion Subjekt im Vorfeld **nicht** gleich groß wie in L1-Texten.

So wie die beiden Hypothesen dastehen, sind sie komplementär gebildet, d.h., wenn die eine zutrifft, gilt die andere nicht und umgekehrt. Die Entscheidung, wann eine der Hypothesen zutrifft, ist allerdings nicht ganz trivial, da man berücksichtigen muss, dass wir

die Häufigkeit ja nur in einer Stichprobe aller möglichen L2-Texte nachzählen können, so dass die beobachteten Häufigkeiten rein zufällig gleich sein oder eben rein zufällig von einander abweichen können. Die große Frage ist nun, wie groß der beobachtete Unterschied sein muss, um die Nullhypothese guten Gewissens ablehnen zu können. Reicht es aus, wenn die Häufigkeit von Subjekten im Vorfeld von L2-Texten um 5 von der in L1-Texten abweicht? Um diese Frage zu beantworten, könnten Sie die absoluten Häufigkeiten von Vorfeldern mit und ohne Subjekt in L2-Texten den absoluten Häufigkeiten von Vorfeldern mit und ohne Subjekt in L1-Texten gegenüberstellen und einen statistischen Signifikanztest durchführen[6]. Siehe Abschnitt 6.5 zur weiterführende Literatur.

6.3 Variablen und ihre Ausprägungen

6.3.1 Urdatenset

Ein wichtiger Schritt bei Korpusstudien wie dem Vergleich der Vorfeldbesetzung bei L2- und L1-Sprechern des Deutschen ist die Erstellung des *Urdatensets*. Im Urdatenset sammelt man detailliert die (Korpus-)Evidenz für jede einzelne Untersuchungsinstanz. Es wird systematisch in der Form einer Tabelle aufgebaut: Jede Zeile steht für eine Untersuchungsinstanz, jede Spalte für ein Merkmal. Tabelle 13 zeigt beispielhaft ein Urdatenset für eine Untersuchung zur Vorfeldbesetzung von fortgeschrittenen chinesischen Deutschlernern. Die Daten stammen aus dem kleinen Lernerkorpus ALeSKo[7], das manuell u.a. mit syntaktischen Funktionen und Phrasenkategorien annotiert wurde. Der Ausschnitt in der Tabelle bezieht sich auf einen Text zum Thema *Ist Urlaub die vergebliche Flucht aus dem Alltag?* Die letzten beiden Spalten beinhalten Informationen zum Vorfeld und dem Rest des jeweiligen Quellsatzes. Das Vorfeld ist die eigentliche Annotationsinstanz. Der Rest des Satzes ist nur informationshalber aufgeführt, um nicht immer zurück zum Korpus gehen zu müssen, wenn man einen Satz nachlesen möchte.

Für das Urdatenset ist es egal, ob die Informationen aus den Annotationen eines Korpus extrahiert wurden oder ob die Beispiele erst nach der Extraktion weiter analysiert wurden. Wichtig ist die Systematik des Tabellenaufbaus: Der Ausschnitt des Urdatensets in Tab. 13 umfasst acht Instanzen: a_1 bis a_8. Jede Instanz wird durch sechs Merkmale charakterisiert. Merkmale sind dann sinnvoll, wenn sie *Variablen* sind, d.h., wenn sie mehr als nur eine *Ausprägung* (auch *Wert*) annehmen können. Würde der Ausschnitt die gesamte Datenmenge repräsentieren, dann wäre das Merkmal *Datei* keine Variable, da es hier nur die Ausprägung *wdt07_01* annimmt. Die etwas kryptisch wirkende Dateibezeichnung leitet sich übrigens vom Studiengang „Wirtschaftssprache Deutsch und Tourismusmanagement" (wdt) ab, in dessen Rahmen der Aufsatz im Jahr 2007 geschrieben wurde. Da das eigentliche Korpus aber mehr als nur einen Text umfasst, ist das Merkmal *Datei* tatsächlich eine Variable mit den Werten: wdt07_01, wdt07_02, ..., wdt07_25 (für die 25 Texte, die 2007 erhoben wurden), wdt08_01, ... wdt08_18 (für die 18 Texte,

---

[6] Bei Häufigkeiten kommt z.B. der χ^2-Test (sprich „Chi-Quadrat"-Test) in Frage, s. z.B. Meindl (2011), Kap. 9.3.; online zu rechnen z.B. auf `http://vassarstats.net/tab2x2.html`. Besser wäre es allerdings, über Textausschnitte Mittelwerte für L2 und L1 zu berechnen.

[7] Vgl. Zinsmeister und Breckle (2012).

Datei	ID	Funktion	Kategorie	Vorfeld	Rest
wdt07_01	a_1	Subjekt	NP	Das	ist doch schwer zu sagen.
wdt07_01	a_2	Adverbial	AP	Einerseite	reisen die Leute, weil sie vom Alltag flüchten möchten.
wdt07_01	a_3	Subjekt	NP	Ihre Arbeit	sind langweilig oder stressig.
wdt07_01	a_4	Adverbial	AP	deswegen	fahren sie irgendwohin, um ihre Ruhe zu bekommen.
wdt07_01	a_5	Adverbial	Satz	Anderseite wenn wir die Touristen genauer beo bachten,	ist es nicht schwer zu erkennen, dass meiste Touristen reich und Priviliger aus der Gesellschaft sind.
wdt07_01	a_6	Prädikativ	na	Was solche Leute benötigen,	ist, andere Länder kennenzulernen, andere Kultur zu verstehen.
wdt07_01	a_7	Subjekt	NP	Sie	sind gar keine Flüchter aus dem Alltag.
wdt07_01	a_8	Adverbial	AP	Deshalb	bin ich felsenfest davon überzeugt, dass Urlaub nicht die vergebliche Flucht aus dem Alltag ist.

Tabelle 13: Ausschnitt aus einem Urdatenset zur Untersuchung der Vorfeldbesetzung im ALeSKo-Korpus

die 2008 erhoben wurden) und zusätzlich Kennungen für L1-Vergleichstexte. Tabelle 14 fasst die Variablen und ihre Ausprägungen zusammen.

Variable	Ausprägungen
Datei	wdt_01, (wdt_02 ...)
ID	a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, (a_9 ...)
Funktion	Adverbial, Prädikativ, Subjekt, na, (Akkusativobjekt, Dativobjekt, es-Korrelat, Präpositionalobjekt, andere)
Kategorie	AP, NP, Satz, (Infinitiv, PP, andere)
Vorfeld	„Das", „Einerseite", ...
Rest	„ist doch schwer zu sagen.", „reisen die Leute, weil sie vom Alltag flüchten möchten.", ...

Tabelle 14: Zur Untersuchung der Vorfeldbesetzung im ALeSKo-Korpus: Variablen aus Tabelle 13 und ihre Ausprägungen

Ausprägungen, die im Ausschnitt in Tab. 13 nicht vorgekommen sind, werden in Klammern angegeben. Eine Besonderheit ist die Ausprägung *na* der Variable *Kategorie* von Instanz *a_6*. *na* steht für „nicht anwendbar" bzw. „nicht verfügbar" (engl. ‚not applicable' / ‚not available'). Tatsächlich handelt es sich hierbei um einen Fehler in den Daten,

eine vergessene Annotation. Die Vorfeldkonstituente von a.6 ist ein freier Relativsatz, der nach den ALeSKo-Guidelines eigentlich die Kategorie *Satz* erhalten sollte[3]. Dass die Kategorien *Vorfeld* und *Rest* auch Variablen sind, mag erst einmal überraschen, trifft aber im formalen Sinn zu. Denken Sie einmal kurz darüber nach.

6.3.2 Skalenniveaus

Das Konzept der Variable ist für viele quantitative Untersuchungen grundlegend. Variablen werden nach der Art ihrer Ausprägungen unterschieden. Was damit gemeint ist, kann man sich am besten anhand von Beispielen verdeutlichen. Wir betrachten hierfür drei verschiedene Studien, bei denen jeweils eine Variable untersucht wird, und geben jeweils die Werte der ersten fünf Instanzen an.

- Studie 1 „Vorfeld":
  Variable: Funktion der Vorfeldkonstituente (vgl. Tab. 13)
  Instanzen 1–5: Subjekt, Adverbial, Subjekt, Adverbial, Adverbial
- Studie 2 „Vokabeln":
  Variable: Vokabelschwierigkeit
  Instanzen 1–5: leicht, leicht, schwer, leicht, mittel
- Studie 3 „Alter":
  Variable: Alter von Teilnehmern einer Studie (vgl. Tab. 16)
  Instanzen 1–5: 22, 20, 25, 18, 29

Wie unterscheiden sich diese Variablen? Wenn Sie zwei beliebige Instanzen einer Variable betrachten, können Sie in allen drei Studien feststellen, ob die beiden Ausprägungen gleich oder verschieden sind: *Subjekt* ist verschieden von *Adverbial*; *leicht* ist gleich *leicht*; *22* ist verschieden von *20*. Sie können also Kategorien unterscheiden und die Instanzen entsprechend auszählen. Als nächstes versuchen Sie, die Instanzen nach den Ausprägungen zu ordnen: *Leicht* ist niedriger auf der Skala *Vokabelschwierigkeit* als *mittel*; *22* ist ein höheres Alter als *20*. Die Ausprägungen *Subjekt* und *Adverbial* zu ordnen ist hingegen nicht möglich. Eine Ordnung nach Alphabet oder Wortlänge würde sich auf die Wörter selbst nicht aber auf die Ausprägungen der Variable *Funktion* beziehen. Als drittes versuchen Sie die Differenzen zu quantifizieren: *22* ist um zwei größer als *20*. Es ist nicht möglich die Differenz zwischen *leicht* und *mittel* auf die gleiche Art und Weise zu quantifizieren. Anders als bei den Ausprägungen der Variable *Alter* ist hier das Intervall zwischen den Werten nicht weiter strukturiert. Es gibt keine messbaren Einheiten. Zuletzt versuchen Sie, Verhältnisse zwischen den Ausprägungen zu quantifizieren: Die Hälfte von *22* ist zwar kleiner als *20*, aber immer noch größer als die Hälfte von *20*. Versuchen Sie einmal die Hälfte der Ausprägung *mittel* zu bestimmen. Und wie sieht es mit der Ausprägung *Adverbial* aus?

Wir fassen zusammen, dass sich die Variablen in ihren Rechenmöglichkeiten unterscheiden. Man spricht hier von verschiedenen *Skalen*. In Tab. 15 sind die drei wichtigsten Skalentypen für korpuslinguistische Studien aufgeführt: die *Nominal-*, *Ordinal-* und *Verhältnisskala* (auch *Ratioskala* vgl. englisch ‚ratio'). Wobei nur die Verhältnisskala eine sogenannte *metrische* Skala ist, mit der Sie im klassischen Sinne rechnen können.

[3] A.5 ist den Guidelines gemäß als Satz annotiert, auch wenn der Ausdruck komplex ist.

Typ	Skala	Merkmal	Beispiele
Nicht-metrisch	1. Nominal	Kategorien Klassenbildung	Syntaktische Funktion (Subjekt, Objekt, ...) Teilnehmer-IDs (P1, P2, ...)
	2. Ordinal	Rangordnung	Vokabelschwierigkeit (leicht, mittel, ...) Bewertungsskalen (z.B. 1–7) Schulnoten
Metrisch	3. Verhältnis (Ratio)	Messbare Intervalle ausgehend von einem Nullpunkt Vergleich von Differenzen und Quotienten	Alter Satzlänge Reaktionszeit Anzahl von Subjekten im Vorfeld (pro Textausschnitt)

Tabelle 15: Die drei wichtigsten Skalentypen für die Korpuslinguistik

Wenn Sie Tab. 15 genauer betrachten, wundern Sie sich vielleicht, dass Schulnoten und Bewertungsskalen, wie z.B. die Likert-Skala, die Sie vielleicht aus der Psycholinguistik kennen, nur ordinalskaliert sein sollen. Ordinal bedeutet schließlich, dass die Intervalle zwischen den einzelnen Werten nicht weiter definiert sind, so dass man z.B. nicht addieren kann. Nichtsdestotrotz werden im Alltag für Schulnoten und Bewertungsskalen Mittelwerte gebildet, die man durch Addieren und Teilen durch die Gesamtzahl ermittelt. Die gelebte Praxis weicht hier von den mathematischen Gegebenheiten ab. Allerdings wird der Skalenstatus für Beispiele dieser Art unter Mathematikern kontrovers diskutiert[9]. Manche argumentieren dafür, die genannten Beispiele als sogenannte intervallskalierte Daten zu interpretieren. Diesen Skalentyp, die *Intervall-Skala*, haben wir in der Tabelle nicht aufgeführt. Es handelt sich hierbei um eine weitere metrische Skala, was bedeutet, dass die Intervalle zwischen den einzelnen Werten wohl definiert sind und dass man sinnvoll Differenzen, Summen und auch Mittelwerte bilden kann. Der Unterschied zwischen einer Intervallskala und einer Verhältnisskala ist, dass Intervallskalen keinen natürlichen Nullpunkt besitzen, Verhältnisskalen aber schon. Was das bedeutet, wird oft an zwei unterschiedlichen Skalen für die Angabe von Temperatur verdeutlicht: die Kelvin-Skala und die Celsius-Skala.

Das Besondere an der Celsius-Skala ist, dass die Werte von −273,15 °C (dem absoluten Nullpunkt) über 0 °C bis in den positiven Bereich gehen. Da der Nullpunkt zwar physikalisch motiviert (Gefrierpunkt von Wasser), aber letztendlich mathematisch beliebig gewählt ist, sind die Ausprägungen der Variable *Celsius-Skala* nur intervall- aber nicht verhältnisskaliert. Sie können zwar feststellen, dass 22 °C um 2 °C wärmer ist als 20 °C. Sie können aber nicht wirklich behaupten, dass 10 °C nur halb so warm ist wie 20 °C. Wenn doch, versuchen Sie einmal das Verhältnis zwischen 5 °C und −5 °C zu beschreiben.

[9] Vgl. die Hinweise in Bortz und Schuster (2010), S. 22–23.

Anders als die Celsius-Skala ist die Kelvin-Skala tatsächlich verhältnisskaliert, da ihr Skalen-Nullpunkt mit dem absoluten Temperaturnullpunkt übereinstimmt.

Wenn man die Skalentypen wie in Tab. 15 von oben nach unten ordnet, wird deutlich, dass alle Operationen, die mit einem ‚niedrigeren' Skalentyp möglich sind, auch mit Vertretern der höheren Skalentypen durchführbar sind. Man kann also verhältnisskalierte Zahlwerte immer auch in eine Rangordnung bringen, man verzichtet dabei allerdings auf Information, da eine Rangordnung nichts über die Differenzen zwischen den Rängen aussagt.

Sie müssen sich immer über die Skalenniveaus ihrer Variablen im Klaren sein, da diese die mathematischen Möglichkeiten im Umgang mit den Variablen beschränken. Dabei dürfen Sie sich nicht von der äußeren Form der Ausprägungen täuschen lassen. Die Variable *Teilnehmer* z.B. ist nominalskaliert. Wir können die Ausprägungen *Claus, Max, Jonas, Anne* usw. lediglich auszählen (*Wie viele Instanzen beobachten wir z.B. für die Ausprägung Claus?*). Wenn Sie nun für die Teilnehmer an der Studie anstelle von Namen Identifikationsnummern notieren (z.B. die Matrikelnummer bei Studierenden), dann erhält die Variable *Teilnehmer* zwar Ausprägungen wie *553483, 283111* usw., bleibt aber weiterhin nominalskaliert. Das heißt, sie dürfen weiterhin nur auszählen, aber nicht rechnen. Es wäre evtl. zu prüfen, ob die Matrikelnummern eine interne Rangordnung symbolisieren (z.B. *Eine niedrigere Zahl entspricht einer längeren Studiendauer*). Im positiven Fall wäre die Variable dann ordinalskaliert.

6.4 Zwei Auswertungsbeispiele

Nachdem oben das ALeSKo-Korpus die Grundlage für das Beispiel zur Erstellung eines Urdatensets bildete, wendet sich dieser Abschnitt einem anderen Lernerkorpus des Deutschen zu, welches den Charme hat, dass es Texte von Lernern mit drei unterschiedlichen Erstsprachen umfasst: Das Kobalt-Korpus beinhaltet insgesamt 77 Texte von chinesischen, weißrussischen und schwedischen Deutschlernern sowie parallel erhobene Texte deutscher Muttersprachler. Alle Texte wurden an Universitäten oder Gymnasien in den Herkunftsländern unter kontrollierten Umständen erhoben[10].

Das Kobalt-Korpus (v1.6) besteht aus vier Subkorpora, die nach den Erstsprachen der Teilnehmer benannt sind[11]:

- BEL: 19 Texte von weißrussischen (engl. ‚Belarus') bzw. russischen Studierenden
- CMN: 20 Texte chinesischer Studierender, die Mandarin als Erstsprache haben
- SWE: 18 Texte schwedischer Studierender und Gymnasiasten
- DEU: 20 Texte deutscher Gymnasiasten als Kontrolltexte

Die beiden Auswertungsbeispiele sollen illustrieren, wie man unterschiedliche Variablentypen beschreiben, darstellen und testen kann. Am Ende des Abschnitts werden die Variablentypen in einer kleinen Übersicht zusammengefasst.

---

[10] Vgl. `www.kobalt.de`.

[11] Die Benennung folgt der Drei-Buchstaben-Kennung für Sprachnamen nach ISO 639-3.

6.4.1 Alter von L2-Lernern

Beim ersten Auswertungsbeispiel werfen wir einen genaueren Blick auf die Metadaten des Kobalt-Korpus. Die Metadaten dokumentieren die Erhebungsumstände und geben Hintergrundinformationen zu den Teilnehmern in der Form von ausführlichen Lernbiographien[12]. Eine grundlegende Information ist dabei das Geburtsjahr der Probanden. Für unser Beispiel wurde aus der Differenz zwischen Geburtsjahr und Erhebungsdatum eine zusätzliche Variable *Alter* abgeleitet. Ziel dieser kleinen Studie ist es nun, das Alter der Probanden, d.h. die Variable *Alter*, in den vier Subkorpora zu vergleichen. Hierzu muss man sich zunächst eine Vorstellung von der Werteverteilung innerhalb der einzelnen Subkorpora verschaffen. Die Spalte *Alter* im Urdatenset, s. Tab. 16, liefert diese Information für die Texte 1 bis 10 des schwedischen Subkorpus SWE.

Identifikator	Subkorpus	Alter
Kobalt_SWE_001_2011_06	SWE	22
Kobalt_SWE_002_2011_05	SWE	20
Kobalt_SWE_003_2011_05	SWE	25
Kobalt_SWE_004_2011_11	SWE	18
Kobalt_SWE_005_2011_12	SWE	29
Kobalt_SWE_006_2011_12	SWE	19
Kobalt_SWE_007_2011_12	SWE	21
Kobalt_SWE_008_2011_12	SWE	20
Kobalt_SWE_009_2012_05	SWE	33
Kobalt_SWE_010_2012_05	SWE	25
...	...	

Tabelle 16: Kobalt-Metadaten: Alter in Jahren (Ausschnitt)

Schon dieser kleine Datenausschnitt macht ein grundlegendes Problem deutlich: Unsortierte Zahlenangaben sind sehr schwer zu verstehen. Außerdem irritiert hier die komplexe Identifikatorangabe in der ersten Spalte, die bei der Erfassung der Altersverteilung keine weitere Rolle spielt. Tabelle 17 ist aus Tabelle 16 abgeleitet, indem die Werte der Variable *Alter* aufsteigend sortiert wurden. Der Übersicht halber wurde die Tabelle zusätzlich auf die unmittelbar relevanten Spalten reduziert. Aus der Spalte *Alter* kann man nun ablesen, dass die Werte zwischen 18 und 33 liegen, dass 20 und 25 je zweimal vorkommen usw. Aufgrund der neuen Anordnung ist es viel einfacher, sich eine konkrete Vorstellung von der Werteverteilung zu machen.

Noch besser kann man sich metrische Datensätze, d.h. Zahlenreihen, vorstellen, wenn sie grafisch dargestellt werden. *Streudiagramme* sind eine einfache Möglichkeit, dies zu tun. Je nachdem, mit welcher Software man die Grafik erzeugt, müssen die Daten gegebenenfalls zuerst in eine kompakte Darstellung überführt werden, aus der die Grafik dann generiert werden kann. Tabelle 18 ist eine solche kompakte Auflistung der

[12] Das Kobalt-Korpus orientiert sich hier am Falko-Korpus vgl. Reznicek et al. (2012).

Subkorpus	Alter
SWE	18
SWE	19
SWE	20
SWE	20
SWE	21
SWE	22
SWE	25
SWE	25
SWE	29
SWE	33
...	...

Tabelle 17: Kobalt-Metadaten: Alter in Jahren – sortiert (Ausschnitt)

Werte von *Alter* für alle vier Kobalt-Subkorpora. Hier ist für jede Ausprägung von *Subkorpus* eine Spalte angelegt, welche die mit dieser Ausprägung auftretenden Werte von *Alter* enthält.

Abbildung 20 zeigt ein Streudiagramm: Die vertikale y-Achse erfasst das Alter in Lebensjahren; die horizontale x-Achse stellt einen simplen Index dar, über den die einzelnen Einträge angeordnet werden. Beachten Sie, dass die y-Achse nicht bei Null beginnt, sondern dem Wertebereich angepasst ist, den die Daten abdecken. Zusätzlich wurden zwischen den einzelnen Datenpunkten eines Subkorpus Linien gezogen, so dass man die Verteilungen besser erfassen kann. Hierbei ist wichtig, dass Sie sich klar machen, dass die Linien reine Hilfskonstrukte sind. Die Indexzahlen repräsentieren Teilnehmer-Instanzen und sind daher nicht als Zahlen zu interpretieren, sondern als nominale Werte. Genauso gut hätten hier Platzhalter wie *P1*, *P2* usw. stehen können oder, wenn nur eines der Subkorpora abgebildet worden wäre, sogar die Namen der Probanden. Die Zwischenräume zwischen den Werten auf der x-Achse haben keinerlei Bedeutung. Auch bei den Altersangaben auf der y-Achse handelt es sich um *diskrete* Werte, nicht um ein Kontinuum, das den Verbindungslinien entsprechen würde, da immer nur ganze Zahlen als Altersangabe eingetragen sind.

Um eine *metrische* Variable wie die Altersangabe genauer zu charakterisieren verwendet man *Kennwerte*. Häufig verwendet werden Kennwerte für die *zentrale Tendenz* oder *Lage* und für die *Streuung* der Datenpunkte: Der *Durchschnitt* bzw. das *arithmetische Mittel* („Mittelwert") $\bar{x}$ berücksichtigt den gesamten Werteraum einer Variable, da es aus der Summe aller Werte geteilt durch die Anzahl der Werte errechnet wird. Für den Altersdurchschnitt im schwedischen Subkorpus (SWE) erhält man zum Beispiel das (gerundete) Mittel:[13]

$$\bar{x}_{\text{SWE}}(Alter) = \frac{18 + 18 + 18 + 19 + \cdots + 25 + 29 + 33 + 54}{18} \approx 23{,}8$$

[13] Für die fehlenden Werte siehe Tab. 18.

BEL	CMN	SWE	DEU
20	20	18	18
21	20	18	18
21	20	18	19
21	20	19	19
21	21	19	19
22	21	19	19
22	21	20	19
22	21	20	20
22	21	20	20
22	22	21	20
22	22	22	20
22	22	23	20
22	22	25	20
22	22	25	20
22	22	25	20
22	22	29	20
23	22	33	20
23	23	54	21
26	23		21
	24		21

Tabelle 18: Vollständige Altersverteilung im Kobalt-Korpus (Mögliche Eingabe für eine Diagrammerstellung)

Als Streuung wird oft die *Standardabweichung* s (auch *sd* von engl. *standard deviation*) angegeben. Sie gibt die durchschnittliche Abweichung aller Werte vom errechneten Mittelwert an und ignoriert dabei, ob die einzelnen Abweichungen positiv oder negativ sind. In der Berechnung wird letzteres durch Quadrieren erreicht („hoch zwei"). Um das Ergebnis dem ursprünglichen Zahlenraum anzupassen, wird am Ende die Wurzel gezogen (hier mit gerundeten Werten):

$$s_{\mathrm{SWE}}(Alter) = \sqrt{\frac{(18-23{,}8)^2+(18-23{,}8)^2+\cdots+(33-23{,}8)^2+(54-23{,}8)^2}{18}}$$
$$= \sqrt{\frac{33{,}64+33{,}64+\cdots+84{,}64+912{,}04}{18}} \approx 8{,}3$$

! Beachten Sie, dass es zweierlei Berechnungswege für die Standardabweichung gibt, je nachdem, ob man eine bekannte Verteilung beschreibt oder mittels einer Stichprobe eine unbekannte Grundgesamtheit schätzt. Handelt es sich nur um eine Stichprobe, wird nicht durch die beobachtete Anzahl von Instanzen n geteilt, sondern durch $n-1$,

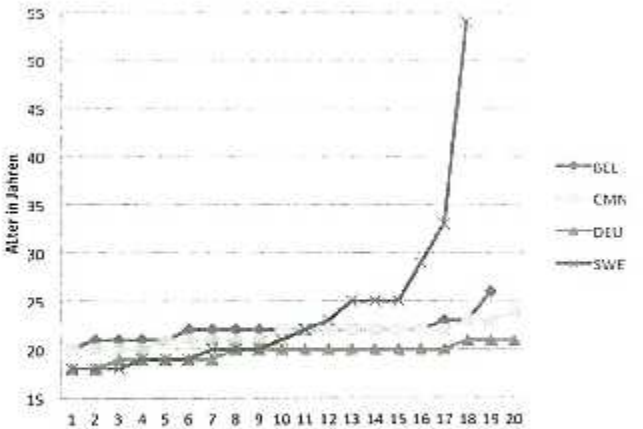


Abbildung 20: Streudiagramm der Altersverteilung im Kobalt-Korpus (mit MS Excel erstellt)

was einen etwas größeren Wert für die Standardabweichung ergibt und damit für den Schätzwert realistischer ist, weil ja auch der beobachtete Mittelwert der Stichprobe mit großer Wahrscheinlichkeit vom echten Mittelwert der Grundgesamtheit abweicht. Beachten Sie, dass Statistikprogramme im Zweifelsfall die Standardabweichung für Stichproben errechnen.

Zusammenfassend ergibt sich für die Verteilung der Variable *Alter* im schwedischen Subkorpus SWE der Mittelwert 23,8 und die Standardabweichung 8,3 (jeweils gerundet). Kürzer ausgedrückt: $\bar{x} = 23,8$, $s = 8.3$ (auch: $\varnothing = 23,8 \pm 8.3$). In Tab. 19 sind die Mittelwerte und Standardabweichungen für die vier Subkorpora zusammengefasst dargestellt (mit gerundeten Werten).

Subkorpus	Mittelwert	Standard-abweichung	Minimum	Maximum	Median
BEL	22,0	1,2	20	26	22,0
CMN	21,6	1,1	20	24	22,0
SWE	23,8	8,3	18	54	20,5
DEU	19,7	0,8	18	21	20,0

Tabelle 19: Zentrale Tendenzen und Streuung der Variable *Alter* (gerundet)

Demnach sind die schwedischen Teilnehmer mit 23,8 Jahren im Durchschnitt am ältesten, die deutschen Teilnehmer mit durchschnittlich 19,7 Jahren am jüngsten. Betrach-

tet man die Standardabweichungen, fällt ins Auge, dass das Alter im schwedischen Subkorpus wesentlich weiter streut als in den anderen Subkorpora. Im Streudiagramm in Abb. 20 konnte man schön sehen, dass der Großteil der schwedischen Lerner um die 20 Jahre alt ist und dass lediglich eine kleine Anzahl von *Ausreißern* wesentlich älter ist. Um zu verhindern, dass solche einzelnen, extremen Werte, den Blick auf die breite Masse der Datenpunkte verzerren, empfiehlt es sich die Extremwerte (*Minimum, Maximum*) der Verteilung anzuschauen. Daran angelehnt ist auch ein weiteres Maß für die zentrale Tendenz, der Median, der anders als der Mittelwert nicht unbedingt Thema des Mathematikunterrichts in der Schule ist: Der *Median* entspricht dem Wert des Datenpunkts, der sich genau zwischen dem minimalen und dem maximalen Datenpunkt befindet, wenn man bei einer ungeraden Instanzenanzahl alle Werte der Größe nach sortiert hat (vgl. BEL in Abb. 21). Ist die Anzahl der Datenpunkte gerade, berechnet man den Median als den Mittelwert der beiden mittleren Datenpunkte (vgl. die Zahl in Klammern bei CMN, SWE, DEU in Abb. 21). Ein Vergleich der Mediane mit den Mittelwerten in Tab. 19 zeigt, dass der Median robust gegenüber Ausreißern ist. Vergleicht man die Variable *Alter* in den vier Subkorpora anhand des Medians, dann ist die schwedische Lernergruppe mit 20,5 Jahren jünger als die weißrussische und die chinesische (jeweils 22,0 Jahre).

		Median	
BEL	20 21 21 21 21 22 22 22 22 22	22	22 22 22 22 22 22 23 23 26
CMN	20 20 20 20 21 21 21 21 21 22	(22)	22 22 22 22 22 22 22 23 23 24
SWE	18 18 18 19 19 19 20 20 20	(20,5)	21 22 23 25 25 25 29 33 54
DEU	18 18 19 19 19 19 19 20 20 20	(20)	20 20 20 20 20 20 20 21 21 21
Quantil:	0%	50%	100%

Abbildung 21: Der Median (= 50%-Quantil) als zentrale Tendenz

Der Ausdruck *Quantil* in Abb. 21 bezeichnet ein sogenanntes Lagemaß: Ein x%-Quantil ist ein Schwellwert, für den gilt, dass x Prozent aller Werte kleiner sind als er (oder gleich). Die 0%- und 100%-Quantilen entsprechen dem Minimum und dem Maximum. Das 50%-Quantil ist der *Median*: Er besagt, dass 50% aller Werte kleiner oder zumindest gleich sind wie der Wert, den er selbst annimmt.

Die Beschreibung einer Verteilung durch Quantile kann sehr informativ sein, vgl. Tab. 20: Zwischen dem 25%- und 75%-Quantil liegen die Hälfte aller Datenpunkte. Dieser Wertebereich ist also sehr charakteristisch für die jeweilige Verteilung.
Weiter oben hatten wir die geordnete Wertetabelle durch ein Streudiagramm grafisch veranschaulicht. Für die zusammenfassende Darstellung auf der Basis von Quantilen existiert ebenfalls eine Visualisierungsmöglichkeit: Der *Boxplot* (auch: Kasten-Grafik), vgl. Abb. 22: Der Kasten (engl. ‚box') markiert den oben bereits genannten Bereich zwischen dem 25%- und 75%-Quantil und beherbergt damit 50% der Datenpunkte. Der dicke Strich steht für den Median (50%-Quantil). Die Differenz zwischen dem 25%- und 75%-Quantil, d.h. die Länge des Kastens, wird als *Interquartilsabstand* bezeichnet. Dieser wird oft dafür verwendet, die Ausdehnung der *Antennen* (engl. ‚whiskers')[14] zu

[14] Engl. für Katzenschnurrhaare.

	0%	25%	Median	Mittelw.	75%	100%
BEL	20,0	21,5	22,0	22,0	22,0	26,0
CMN	20,0	21,0	22,0	21,6	22,0	24,0
SWE	18,0	19,0	20,5	23,8	25,0	54,0
DEU	18,0	19,0	20,0	19,7	20,0	21,0

Tabelle 20: Zusammenfassende Darstellung der Werteverteilungen durch Quantile und arithmetisches Mittel

berechnen, die dazu dienen bei den Datenpunkten, die außerhalb des Kastens liegen (immerhin 50% aller Werte), zwischen regulären Werten und *Ausreißern* zu unterscheiden. Eine Variante, die man häufig antrifft, ist die, dass die Antennen maximal eine Länge von 1,5 Interquartilsabständen einnehmen. Falls die Werte nicht so weit streuen, sind die Antennen entsprechend kürzer. Alle Datenpunkte, deren Werte außerhalb der Antennen liegen, können in Bezug auf die restliche Datenverteilung als Ausreißer betrachtet werden. Die kleinen Kreuze stellen die Mittelwerte dar. Sie wurden hier der Vollständigkeit halber ergänzt, sind aber nicht zwingend Bestandteil eines Boxplots.

Die Boxplots in Abb. 22 zeigen, dass sich die Verteilungen des Alters in BEL und CMN sehr ähneln. Insgesamt ist die Masse der Werte in BEL am kompaktesten angeordnet, auch wenn es ein paar Ausreißer gibt, die schlussendlich für die etwas größere Standardabweichung im Vergleich zu CMN verantwortlich sind (vgl. Tab. 19). Der lange Kasten bei SWE verdeutlicht schön, dass das Alter in SWE insgesamt weit streut. Der eine, ganz extreme Ausreißer hat sicher einen verzerrenden Einfluss auf das arithmetische Mittel von SWE (siehe das Kreuzchen). Für den Vergleich des Alters in den Subkorpora kann man anhand der Boxplots zusammenfassen, dass sich BEL und CMN kaum unterscheiden. SWE weicht in Bezug auf Mittelwert und Median von BEL und CMN ab (einmal nach oben und einmal nach unten) und weist eine wesentlich größere Streuung auf. Das Alter von DEU ist im Durchschnitt niedriger als bei den anderen Subkorpora, sowohl in Bezug auf den Mittelwert als auch den Median. Die Gruppe ist recht homogen, die Werte weisen nur eine geringe Streuung auf.

Bekannter als die Darstellung des Medians und der anderen Quantilen im Boxplot ist die Abbildung von Mittelwerten als Balken- oder Säulendiagramme ergänzt um einen *Fehlerbalken*, der die Standardabweichung visualisiert, vgl. Abb. 23. Diese Visualisierung ergibt ebenfalls einen Eindruck von der Streuung der Daten, auch wenn sie weniger informativ ist als die Boxplots: Die Mittelwerte von BEL und CMN sind zwar nicht identisch, aber die Fehlerbalken überlappen fast vollständig. Der Fehlerbalken von SWE hat eine große Ausdehnung und signalisiert, dass die Werte sehr weit um den Mittelwert streuen, was bedeutet, dass das arithmetische Mittel von SWE die Datenverteilung schlechter charakterisiert als die Mittelwerte der anderen Subkorpora. Das Alter in DEU ist im Durchschnitt niedriger als bei den anderen Subkorpora, wobei der Fehlerbalken von SWE auch hier überlappt.

In diesem Abschnitt haben Sie verschiedene Möglichkeiten kennengelernt, die Ausprägungen einer Variable darzustellen, die einer *metrischen* Skala angehören und damit bestimmte Zahlwerte annehmen. Im nächsten Abschnitt betrachten wir eine Variable,

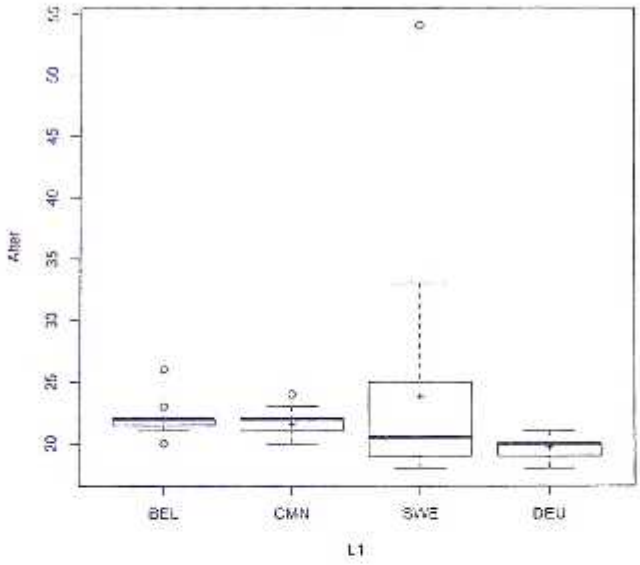


Abbildung 22: Boxplots der Altersverteilung im Kobalt-Korpus (mit R erstellt)

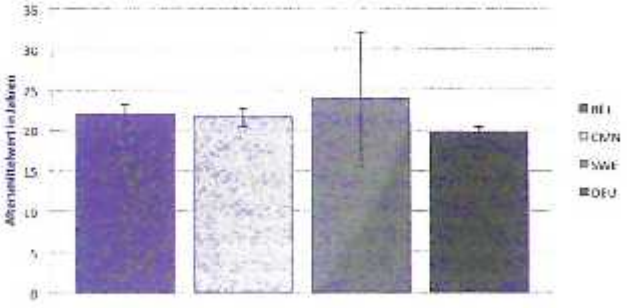


Abbildung 23: Mittelwerte der Altersverteilung im Kobalt-Korpus; Fehlerbalken = ±sd (mit MS Excel erstellt)

deren nominalen Ausprägungen keine Rangfolge aufweisen, sondern lediglich unterschiedlichen Klassen angehören.

6.4.2 L2-Verwendung von Modalverben

Lernersprache weicht auf verschiedenen Ebenen von der Zielsprache ab. Grammatische Fehler wie die falsche Flexionsendung in (7) (*euren* anstatt *eure*) sind offensichtliche Abweichungen.

(7) ...und das ist nicht gut für euren Gesundheit.

Eine subtilere Art der Abweichung besteht im Unter- bzw. Übergebrauch von an sich zielsprachlichen Konstruktionen, die einen Lernertext insgesamt als abweichend erscheinen lassen können. Ein Beispiel hierfür ist die Verwendung von Modalverben wie *können* oder *müssen*. Dieses Beispiel nutzen wir zur Illustration der Darstellung eines nominalen Merkmals. Die zugrundeliegende Untersuchung ist dadurch motiviert, dass Modalität bzw. die Verwendung von Modalverben ein wichtiger Aspekt von argumentativen Texten ist und als Indikator für L2-Kompetenzniveaus diskutiert wird[15].

Die linguistische These, die wir untersuchen wollen, lautet, dass L2-Sprecher Modalverben *anders* verwenden als L1-Sprecher. Wir operationalisieren *anders* zunächst einfach dadurch, dass wir annehmen, dass die Verwendungshäufigkeiten unterschiedlich sind. Wichtig hierbei ist, dass wir nicht einfach zählen können, wie viele Modalverben in den einzelnen Subkorpora auftreten, sondern *relative Häufigkeiten* ermitteln müssen. Die Notwendigkeit hierfür wird schnell klar, wenn man z.B. vergleicht, wie charakteristisch das Auftreten von 100 Modalverben in einem Text mit 200 Sätzen ist im Verhältnis zu 100 Modalverben in einem Text mit 2000 Sätzen. Im ersten Fall tritt durchschnittlich in jedem zweiten Satz ein Modalverb auf. Die relative Häufigkeit von Modalverben in Bezug auf die Satzanzahl ist demnach 0,5. Im zweiten Fall findet man durchschnittlich nur in jedem zwanzigsten Satz ein Modalverb. Hier ist die relativen Häufigkeit nur 0,05. Tabelle 21 fasst die absoluten und relativen Häufigkeiten für dieses kleine Gedankenexperiment unter der Annahme zusammen, dass pro Satz maximal ein Modalverb auftreten kann. Zusätzlich sind Prozentzahlen angegeben, unter denen man sich oft mehr vorstellen kann als unter den relativen Häufigkeiten[16].

Um Vorkommnisse in verschiedenen Datengrundlagen vergleichen zu können, müssen wir demnach die beobachteten, absoluten Häufigkeiten *normalisieren* und dem Vergleich die resultierenden relativen Häufigkeiten zugrunde legen. Dies verlangt, dass wir eine Referenzgröße für die Normalisierung bestimmen. In dem kleinen Beispiel oben wurde als Referenzgröße die Anzahl der Sätze im Text gewählt. Die (stark vereinfachende) Idee dahinter ist, dass man pro Satz maximal ein Modalverb verwendet, so dass die Anzahl der Sätze die Obergrenze für die zu erwartende Häufigkeit der Modalverben vorgibt. Eine andere denkbare Referenzgröße wäre die Anzahl der Token

---

[15] Vgl. Maden-Weinberger (2008).

[16] Bitte beachten Sie, dass Prozentzahlen in der Regel erst ab einer untersuchten Anzahl von 80 angegeben werden, da Prozentzahlen eine Aussage darüber machen, wie groß die Häufigkeit im Schnitt bei 100 untersuchten Einheiten ist. Vielen Dank an Fabian Barteld für diese Klarstellung.

Text	Absolute Häufigkeit Modalverben	Anzahl Sätze	Relative Häufigkeit	Prozentzahl Modalverben
t_1	100	200	$\frac{100}{200} = 0.5$	0,5*100%=50%
t_2	100	2000	$\frac{100}{2000} = 0.05$	0,05*100%=5%

Tabelle 21: Absolute versus relative Häufigkeiten (und Prozentzahlen)

im Text. Bei großen Datenmengen wird auf eine Häufigkeit per 100 000 Token oder per eine Million Token normalisiert. Eine Normalisierung in Bezug auf potenzielle Vorkommenskontexte ist meistens angemessener als eine unspezifische Normalisierung in Bezug auf die gesamte Tokenanzahl – ein Text von 100 Wörtern mit nur vier sehr lange Sätzen, die viele Substantive, Adjektive und Adverbien enthalten, bietet weniger Optionen für die Verwendung von Modalverben als ein gleichlanger Text mit acht relativ kurzen Sätzen. Allerdings ist es nicht immer möglich, die Vorkommenskontexte ohne erheblichen manuellen Aufwand zu bestimmen, wohingegen die Tokenanzahl eine trivial zu ermittelnde Größe ist. Im Zweifelsfall kann man eine Stichprobe in Bezug auf beide Normalisierungsoptionen auswerten, um abzuschätzen, in wie weit die einfachere Option die tatsächlichen relativen Häufigkeiten über- oder unterschätzt.

Für die aktuelle Beispieluntersuchung der Verwendung von Modalverben im Kobalt-Korpus benötigen wir eine einfache, aber sinnvolle Operationalisierung der potenzielle Vorkommenskontexte. Anstelle der Satzanzahl wählen wir die Anzahl an finiten Verben als Bezugsgröße. Die Idee ist, dass jeder finite Teilsatz genau ein finites Verb aufweist, und man so eine adäquatere Referenzgröße erhält, als wenn man nur auf die Anzahl der Sätze Bezug nimmt. Das Kobalt-Korpus ist mit STTS-Wortartentags annotiert, so dass die Häufigkeit der finiten Verben leicht ermittelt werden kann (mittels der STTS-Tags V.FIN)[17]. Beachten Sie, dass diese Operationalisierung ebenfalls vereinfachend ist, da sie zum einen ignoriert, dass Modalverben im Deutschen auch infinit auftreten können, also theoretisch auch in infiniten Teilsätzen möglich sind. Darüber hinaus sind auch Anhäufungen von Modalverben in einem Teilsatz möglich wie *sollte* und *können* in Beispiel (8).

(8) Beides sollte sie verwerten können.

Analog zum ersten Beispiel in Abschnitt 6.2.2 operationalisieren wir die linguistische Fragestellung und leiten daraus eine Hypothese ab:

- *Fragestellung:*
  Unterscheidet sich die Verwendung der Modalverben in L2-Texten von der in L1-Texten?
- *Operationalisierung:*
  L2-Texte enthalten relativ mehr/weniger finite Modalverben (STTS-Tag: VMFIN) als L1-Texte.

[17] „V.FIN" ist eine Abkürzung für VAFIN, VMFIN und VVFIN. Imperative Verben werden hierbei ignoriert.

- *Hypothese:*
  Die relative Häufigkeit von Modalverben (VMFIN) in Bezug auf alle finiten Verben (V.FIN) unterscheidet sich bei L2- und L1-Texten.

Daraus leiten wir eine Nullhypothese H_0 und eine Alternativhypothese H_1 ab.

- *Nullhypothese* H_0: In L2-Texten findet man die gleiche relative Häufigkeit von VMFIN wie in L1-Texten.
- *Alternativhypothese* H_1: In L2-Texten findet man **nicht** die gleiche relative Häufigkeit von VMFIN wie in L1-Texten.

Die Ermittlung der Verhältniszahlen führt wieder zu einem Urdatenset, vgl. z.B. Tab. 13. Neben einem fortlaufenden Index ID enthält es Informationen zum Subkorpus, dem Text, dem STTS-Tag und dem Lemma der Instanz.

ID	Subkorpus	Text	Wortart	Lemma
v1	SWE	001	VVFIN	gehen
v2	SWE	001	VVFIN	fragen
v3	SWE	001	VVFIN	gehen
v4	SWE	001	VAFIN	sein
v5	SWE	001	VVFIN	ankommen
...	...	...	...	...
v11	SWE	001	VMFIN	können
...				

Tabelle 22: Einfaches Urdatenset zur Verteilung von finiten Modalverben im Kobalt-Korpus (Ausschnitt)

Die Variable *Wortart* kann hier die Werte VAFIN, VMFIN, VVFIN annehmen. Es handelt sich hierbei um eine nominale Variable, da ihre Ausprägungen keine Rangordnung (*mehr / weniger von etwas*) implizieren. Sie klassifizieren die Instanzen lediglich in drei unterschiedliche Gruppen. Wertet man die Modalverben in Bezug auf die vier Subkorpora aus, erhält man die Häufigkeitstabelle 23. Die Abkürzung h_n(VMFIN) steht für die relative Häufigkeit von finiten Modalverben, wobei n für die Referenzmenge steht, die hier durch die Anzahl aller finiter Verben (V.FIN) im jeweiligen Subkorpus bestimmt wird.

Grafisch kann man die Häufigkeitsverhältnisse z.B. in *Stapeldiagrammen* darstellen, vgl. Abb. 24. Hierbei lohnt es sich, sowohl die relativen als auch die absoluten Werte zu vergleichen.

Die dunklen Anteile der Balken bei der Prozentdarstellung zeigen von links nach rechts eine leicht sinkende Tendenz. Insgesamt beobachten wir im Subkorpus DEU die geringste relative Häufigkeit von finiten Modalverben in Bezug auf alle finiten Verben, d.h. die L1-Texte unterscheiden sich von den L2-Texten. Ob dieser beobachtete Unterschied groß genug ist, um die Nullhypothese zu verwerfen, die ja besagt, dass kein Unterschied existiert, müsste über weiterführende statistische Tests ermittelt werden.

	BEL	CMN	SWE	DEU	Gesamt
VMFIN	254	172	143	165	734
andere **V.FIN**	1218	955	854	1043	4070
alle V.FIN (=n)	1472	1127	997	1208	4804
h_n(VMFIN)	0,166	0,153	0,143	0,137	0,153

Tabelle 23: Häufigkeiten von finiten Modal- und anderen Verben in den Subkorpora des Kobalt-Korpus (v1.6)

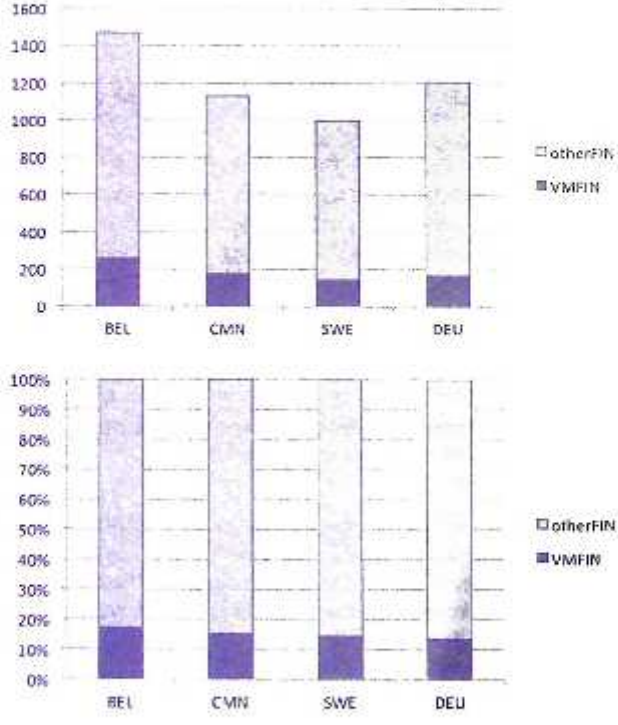


Abbildung 24: Stapeldiagramme mit absoluten (oben) und relativen (unten) Häufigkeiten, letztere in Prozentzahlen dargestellt (mit MS Excel erstellt)

Auf einen wichtigen Punkt müssen wir hier noch hinweisen. Mit der Auswertung in Bezug auf die Subkorpora ignorieren wir, ob nicht ein einzelner Autor enorm viele oder enorm wenige Modalverben zum Korpus beigesteuert hat. Anders ausgedrückt, wir wissen nicht, in wie weit das Gesamtbild durch einen oder wenige Ausreißer verzerrt ist.

Diesem Problem kann Abhilfe geschaffen werden, indem die Auswertung nicht über ein ganzes Subkorpus berechnet wird, sondern über sinnvolle Untereinheiten. In unserem Beispiel bieten sich z.B. die einzelnen Texte als Bezugsgrößen an. Für jedes Subkorpus kann dann ein Mittelwert der relativen Häufigkeiten ermittelt werden und ebenso ein Streuungsmaß wie die Standardabweichung.

6.5 Weiterführende Literatur

Im letzten Kapitel haben wir mehrfach auf weiterführende Erklärungen zur statistischen Berechnung verwiesen. Um den Unterschied zwischen Häufigkeiten oder Mittelwerten statistisch zu prüfen, wendet man Signifikanztests wie den Chi-Quadrat-Test (bei Häufigkeiten) oder den t-Test (bei Mittelwerten) an. Die Herausforderung bei der Anwendung solcher Tests ist, dass jeder Test bestimmte Gegebenheiten voraussetzt, z.B. dass bestimmte Mindesthäufigkeiten vorliegen oder dass sich die Werte einer Verteilung auf eine bestimmte Art verteilen. Um die Anwendung und Interpretation von statistischen Tests zu verstehen, muss man sich von der rein beschreibenden (auch *deskriptiven*) Statistik, wie wir sie im letzten Kapitel angewendet haben, lösen und die erhobenen Daten als Stichprobe einer unbekannten größeren Gesamtheit interpretieren. Eine sehr lesbare deutschsprachige Einführung in Statistik für Sprachwissenschaftler finden Sie bei Meindl (2011). Beim statistischen Testen ist es hilfreich, auf Statistikprogramme zurückgreifen zu können. Gries (2008) ist ebenfalls eine deutschsprachige Statistikeinführung, die gleichzeitig in die kostenlose Statistiksoftware *R* einführt. Ebenfalls auf die Statistiksoftware *R* aufbauend ist die anspruchsvollere englische Einführung von Baayen (2008). Johnson (2008) ist weniger eine systematische (englischsprachige) Einführung, hat aber den Charme, dass die Berechnungen einer Reihe von publizierten Studien Schritt für Schritt in *R* nachgespielt werden. Alle genannten Publikationen verweisen auf Webseiten, von denen die Daten zu den Beispielen vorgehalten werden, so dass man die Berechnungen selbst nachvollziehen kann. Eine weitere, englische Einführung, die aber nicht gleichzeitig in eine Software einführt, ist Oakes (1998). Larson-Hall (2010) bietet eine Schritt-für-Schritt-Einführung in Statistik mit der kommerziellen Software SPSS. Die begleitende Webseite[18] enthält auch eine analoge Einführung mit *R*, die zukünftig ebenfalls als Buch herausgegeben werden soll.

Biber und Jones (2009) geben eine grundlegende Übersicht zum quantitativen Ansatz in der Korpuslinguistik. Evert (2006) entwickelt die Bibliotheks-Metapher, mit der er anschaulich darstellt, wie man Korpusbefunde statistisch interpretieren kann. Perkuhn et al. (2012) diskutieren im Detail, wie man Korpusfrequenzen analysiert. Baroni und Evert (2008) motivieren in einem Handbuchbeitrag Grundlagen der stichprobenbasierten Statistik für die Analyse von Korpusbelegen (z.B. den t-Test). Nicht speziell für Korpuslinguisten eingerichtet, aber sehr brauchbar, ist die Webseite *VassarStats: Website for Statistical Computation* (`http://vassarstats.net/`).

[18] Webseite zu Larson-Hall (2010): `http://cw.routledge.com/textbooks/9780805861853/`.

6.6 Aufgaben

1. Erstellen Sie Urdatensets für die linguistischen Untersuchungen Studie 1 und Studie 2 unten. Machen Sie sich dafür jeweils zunächst bewusst,
   - welche Variablen erhoben werden,
   - mit welchen Merkmalsausprägungen diese Variablen auftreten und welchem Skalentyp die Variablen zuzuordnen sind.
   - Skizzieren Sie dann jeweils ein Urdatenset mit drei fiktiven Instanzen.

   Die folgende Beispielstudie illustriert, wie Sie diese Aufgabe bearbeiten können:

   Beispielstudie Sie wollen die durchschnittliche Satzlänge im „Lenz" von Georg Büchner auf der Basis der Wortanzahl ermitteln. Ein *Wort* bezeichnet hier die orthographische Einheit, die durch Leerstellen oder Satzzeichen von anderen Wörtern getrennt ist (Apostrophe gelten hierbei nicht als Satzzeichen).

   Musterlösung:

   - Variablen: Satz-ID, Satzlänge
   - Ausprägungen von Satz-ID: beliebige Namen, z.B. s_1, s_2, ... (oder auch 1, 2, 3 ...), Skala: Nominal;
     Ausprägungen von Satzlänge: ganze Zahlen, 1, 2, 3 ..., Skala: Verhältnis (metrisch).

Satz-ID	Satzlänge	(Satz)
s_1	7	Den 20. Januar ging Lenz durch's Gebirg.
s_2	17	Die Gipfel und hohen Bergflächen im Schnee, die Thäler hinunter graues Gestein, grüne Flächen, Felsen und Tannen.
s_3	14	Es war naßkalt, das Wasser rieselte die Felsen hinunter und sprang über den Weg.
...	...	...

Tabelle 24: Ausschnitt aus dem Urdatenset zur Beispielstudie: Durchschnittliche Satzlänge in Büchners „Lenz" (Die Angabe des Satzes war nicht explizit gefragt)

Studie 1 Sie wollen die Sprachkenntnis von Fremdsprachlernern untersuchen. Dafür notieren Sie die Komplexität der Sätze, die eine Person im Präsens bilden kann. Sie gehen von drei Komplexitätsstufen aus: (a) einfach: Hauptsätze mit einem Verb, (b) mittel: Hauptsätze mit mehr als einem Verb (z.B. Modalverb und Vollverb) oder (c) komplex: komplexe Sätze (Haupt- und Nebensatz).

Beispielsätze:
(a) *Einfach*: Touristen lieben das Reisen.
(b) *Mittel*: Touristen wollen viel erleben.
(c) *Komplex*: Touristen meinen, dass das Reisen Spaß macht.

Studie 2 Im Florentinischen Italienisch werden Vokale am Wortende oft getilgt. Sie untersuchen, ob diese Tilgung etwas mit dem Typ des Vokals zu tun hat, ob also bestimmte Vokale häufiger als andere getilgt werden[19]. Es kommen dabei folgende Vokale in Frage: [a], [e], [i] und [o]. Die Datenerhebung wird an einem Korpus von gesprochener Sprache vorgenommen. Untersucht wird die Aussprache der folgenden Akkusativpronomen *la, le, li, lo* (sie$_{\text{FemSg}}$, sie$_{\text{FemPl}}$, sie$_{\text{MaskPl}}$, er). Überlegen Sie zuerst, welche Untersuchungseinheiten Sie in dieser Erhebung zu Grunde legen wollen. Bedenken Sie, dass hier Sätze als Basiseinheit nicht sinnvoll sind.

2. Die folgende Zahlenreihe stellt die Satzlängen der ersten elf Sätze eines Lernertexts[20] dar. Für die Satzlängen wurden alle Token mit Ausnahme der Satzzeichen gezählt.
Ihre Aufgabe ist es, die Verteilung der verhältnisskalierten Variable *Satzlaenge* durch Kennwerte zu beschreiben und angemessen grafisch darzustellen.

(9) *Satzlängen eines Lernertexts* (Auszug)
12 10 11 11 11 15 19 11 10 3 19

a. Erstellen Sie eine Tabelle (=Spalte) mit den Werten in einem Tabellenprogramm ihrer Wahl und benennen Sie die Spalte (mit einer Überschrift z.B. *Satzlaenge*).

b. Ordnen Sie die Instanzen der Spalte *Satzlaenge* der Größe nach (aufsteigend). Wie sieht die Verteilung aus? Welche Tendenzen sehen Sie?

c. Erzeugen Sie ein Streudiagramm der Verteilung. Was repräsentieren die beiden Achsen? Benennen Sie diese entsprechend. Bestätigt die Grafik Ihre bisherige Vorstellung der Daten?

d. Berechnen Sie den Mittelwert und die Standardabweichung der Verteilung.

e. Berechnen Sie den 0%-, 25%-, 50%- , 75%- und 100%-Quantilen. Bestimmen Sie das Minimum, das Maximum und den Median.

f. Vergleichen Sie den Mittelwert und den Median. Warum weichen die beiden Kennwerte voneinander ab?

g. Zeichnen Sie einen Boxplot der Verteilung.

[19] Diese Aufgabe ist inspiriert durch Garrapa (2011).
[20] Es handelt sich um den Text Kobalt_SWE_011.

7 Selber kochen oder auswärts essen gehen? — Deutschsprachige Korpora

Wenn Sie dieses Kapitel gelesen haben, dann haben Sie einen Überblick über die Vielfalt deutschsprachiger Korpora. Sie haben eine Korpustypologie kennengelernt, die es Ihnen erlaubt, Korpora systematisch zu klassifizieren. Für Ihre eigenen korpuslinguistischen Projekte bedeutet das, dass Sie hier einen Wegweiser in die Korpuslandschaft bekommen haben, der Ihnen hilft, ein passendes Korpus für Ihr Forschungvorhaben zu finden.

7.1 Einleitung

In diesem Kapitel wollen wir eine Übersicht über die Korpuslandschaft des Deutschen geben. Zum einen möchten wir, dass die wichtigsten Ressourcen für Sie leicht zugänglich sind, und das beginnt damit, dass Sie wissen, welche Ressourcen Sie bei Bedarf konsultieren können. Zum anderen möchten wir es nicht bei einer reinen Aufzählung des Vorhandenen belassen, sondern die Ressourcen in eine korpuslinguistisch begründete Typologie einordnen.

Auf der begleitenden Webseite bieten wir Ihnen eine kommentierte Liste von Korpussammlungen und individuellen Korpora an. Wir verstehen diese Liste als eine von uns verantwortete Auswahl aus größeren Repositorien, wie sie momentan entstehen. Besonders möchten wir hier auf das europäische Projekt CLARIN[1] hinweisen, zu dessen Aufgaben der Aufbau eines *Virtual Language Observatory* gehört[2]. Dort können Sie nach ein- und mehrsprachigen Ressourcen der meisten europäischen Sprachen suchen und sich über die Entwicklung auf dem Laufenden halten.

Das Kapitel ist folgendermaßen aufgebaut. Wir beginnen mit einer Korpustypologie und diskutieren dabei die Kriterien, die wir zur Einordnung der Korpora verwendet haben. In Abschnitt 3 nennen wir Ihnen für jedes Kriterium konkrete Beispielkorpora. Am Ende des Kapitels stellen wir einige neuere, uns besonders interessant erscheinende Korpora und Korpusinitiativen vor und diskutieren in diesem Zusammenhang noch einmal methodische Herausforderungen beim Aufbau und bei der Nutzung dieser Korpora.

---

1 Vgl. www.clarin.eu.

2 Vgl. http://www.clarin.eu/content/virtual-language-observatory.

7.2 Korpustypologie

Damit Sie sich in der Vielzahl der Angaben zurecht finden können, haben wir eine Typologie entworfen, die es erlaubt, die Korpora zu klassifizieren.

Im Folgenden stellen wir die Kriterien vor, nach denen wir die Typologie eingeteilt haben[3]: Funktionalität, Sprachenauswahl, Medium, Annotation, Größe, Persistenz, Sprachbezug, Verfügbarkeit (siehe die Übersicht in Abbildung 25).

Linguistische Annotation und Verfügbarkeit sind für uns wichtig genug, um sie als Kriterien aufzunehmen. Wie schon in Kapitel 2 erläutert wurde, scheiden sich in Bezug auf die (linguistische) Annotation der Primärdaten die Geister. Manch einer plädiert dafür, auf linguistische Annotation in Korpora ganz zu verzichten. Wir sind aber der Meinung, dass die Annotation Teil eines Korpus ist. Das heißt zum Beispiel, dass wir zwischen dem unannotierten *Europarl-Korpus* einerseits und dem *Constraint-Grammar-annotierten Europarl-Korpus* unterscheiden, obwohl beide auf den selben Primärdaten beruhen. Dasselbe gilt auch für das unannotierte *Frankfurter Rundschau Korpus*, die *TIGER-Baumbank* und das *SALSA-Korpus*. Sie beruhen alle auf Daten aus der Frankfurter Rundschau, unterscheiden sich aber durch ihre Annotationsebenen (und Größe).

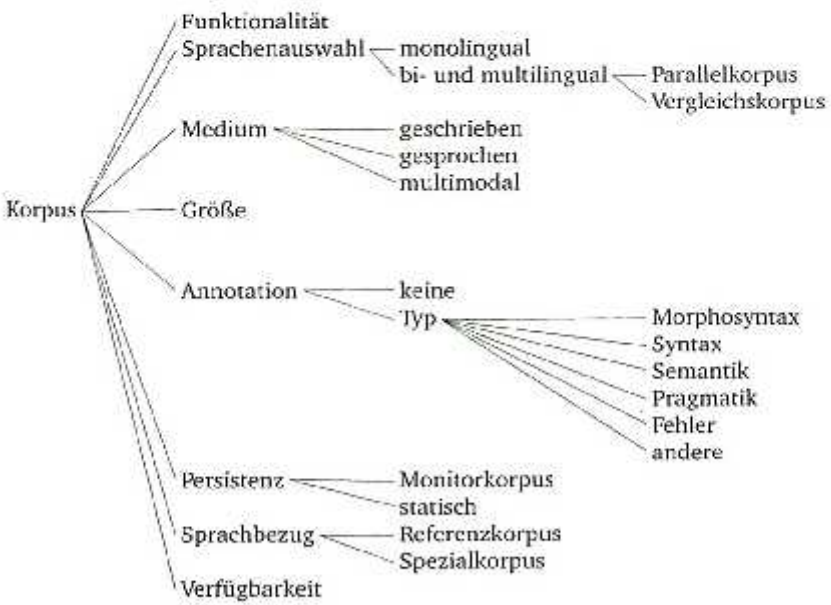


Abbildung 25: Korpustypologie: Übersicht über die Kriterien und ihre Werte

[3] Die Kriterien basieren vorwiegend auf Vorschlägen zum Korpusdesign und zur Korpustypologie, die in Sinclair (1996), Dodd (2000), Kap. 1, Kenny (2000), Engelberg und Lemnitzer (2001), Kap. 1.3, Atkins, Clear und Ostler (1992) und Hunston (2008) vorgestellt und diskutiert werden.

Die Kriterien lassen sich selbst klassifizieren. Zunächst gibt es Kriterien, die die Primärdaten betreffen: Sprachenauswahl, Medium, Größe, Sprachbezug, Funktionalität (Erläuterungen siehe unten). Diese Kriterien sind auch beim Korpusaufbau, also bei der Zusammenstellung der Primärdaten entscheidend. Sie werden als *Designkriterien* (auf Englisch auch als *Sampling*-Kriterien) bezeichnet. Davon zu unterscheiden sind Kriterien, die die Korpusaufbereitung betreffen. In unserer Typologie ist das nur die Annotation. Hier könnte man ggf. weiter unterscheiden, z.B. in positionelle Merkmale, die dem einzelnen Token zugeordnet werden, wie die Morphosyntax, und in strukturelle Merkmale, die potenziell einen wortübergreifenden Charakter haben. Die letzte Gruppe bilden Kriterien, bei denen das physische Korpus im Zentrum steht: Persistenz und Verfügbarkeit.

Kriterium: Funktionalität

Dieses Kriterium bestimmt normalerweise die Festlegung der anderen Kriterien. Zu welchem Zweck wurde ein Korpus ursprünglich erstellt? Der Zweck bestimmt die Designkriterien, den Umfang der Annotation, die Korpusgröße, die Persistenz usw.

> It is a truism that there is no such thing as a ‚good' or a ‚bad' corpus, because how a corpus is designed depends on what kind of corpus it is and how it is going to be used. (Hunston, 2008, S. 155)

Die ursprüngliche Funktionalität erklärt bestimmte Eigenschaften eines Korpus. Sie legt die Nutzung des Korpus aber nicht ein für allemal fest, vgl. die Diskussion um Multifunktionalität in Bezug auf annotierte Korpora, Kapitel 4, S. 60. In Abschnitt 7.3 ab S. 142 stellen wir Ihnen konkrete Beispiele für Funktionalität vor.

Kriterium: Sprachenauswahl

Handelt es sich bei dem dokumentierten Gegenstand des Korpus um eine oder um mehrere Sprachen? Bei *monolingualen Korpora* ist zu beachten, ob innerhalb der Sprache Varietäten erfasst und unterschieden werden, wie etwa beim Deutschen das Schwäbische oder das Schweizerdeutsch. Bei *bilingualen oder multilingualen Korpora* kann man weiter danach unterscheiden, ob es sich

- um *Parallelkorpora* handelt, welche aus Texten in einer Sprache $S1$ und deren Übersetzung(en) in die Sprache(n) $S2 \ldots Sn$ bestehen. Die Textteile, z.B. Absätze oder Sätze, können dabei einander zugeordnet (miteinandern *aligniert*) werden;
- um *Vergleichskorpora* handelt, in welchen Texte mehrerer Sprachen $S1 \ldots Sn$ zu vergleichbaren Diskursbereichen erfasst sind, die aber keine Übersetzungen voneinander sind[4].

[4] In der Übersetzungswissenschaft wird unter *Vergleichskorpus* auch noch ein anderer Korpustyp verstanden. Es handelt sich dabei um ein monolinguales Korpus, das sowohl Texte enthält, die in der Sprache $S1$ originär verfasst wurden, als auch Texte, die von anderen Sprachen nach $S1$ übersetzt wurden. Der ‚originäre' Teil des Korpus dient als Hintergrund, vor dem Besonderheiten von Übersetzungen beobachtet werden können.

Diachrone Korpora, d.h. Korpora, die verschiedene Entwicklungsstufen derselben Sprache dokumentieren, betrachtet man normalerweise als monolingual. Als Zusatzinformation geben wir auch die *Entstehungszeit der Primärdaten* an, was bei einer feineren Untergliederung ein eigenständiges Kriterium wäre.

Kriterium: Medium

Gemeint ist hier das Medium, in dem die Primärdaten entstanden sind. Zu unterscheiden sind Korpora *geschriebener Sprache* von solchen *gesprochener Sprache* und *multimodalen* Korpora, wie z.B. Videokorpora. Bei den gesprochenen Korpora muss man zwischen den Sprachsignalen selbst und den Transkripten unterscheiden. Wir folgen aber Sinclair[5] darin, dass wir auch ein transkribiertes Korpus als Vertreter der *gesprochenen Sprache* zählen. Sinclair weist darauf hin, dass die Grenze zwischen geschriebenen und gesprochenen Texten durchaus unscharf sein kann. Eine geschriebene Rede wurde fürs mündliche Medium konzipiert, ebenso Hörspieltexte und Theaterstücke. Diesen Unterschied haben Koch und Oesterreicher[6] als einen Unterschied zwischen *medialer* und *konzeptioneller* Mündlichkeit bzw. Schriftlichkeit definiert. Man sollte beide Ebenen bei korpuslinguistischen Untersuchungen sorgfältig trennen.

Korpora gesprochener Sprache bestehen manchmal aus vorgegebenen Textmustern, die von professionellen Sprechern eingesprochen werden. Solche Korpora sind für die Sprachgenerierung relevant. Eine andere Mischform zwischen gesprochener und geschriebener Sprache sind Korpora von Chatsprache. Hier wird im schriftlichen Medium konzeptionelle Mündlichkeit realisiert[7]. Wir beschränken uns in dieser Typologie auf die *mediale* Schriftlichkeit bzw. Mündlichkeit.

Bei *multimodalen* Korpora umfassen die Primärdaten weitere Medien wie z.B. eine Videospur, die auch optische Information liefert. Essenziell ist dies zum Beispiel für die Gestenforschung oder für Korpora der Gebärdensprache.

Kriterium: Annotation

Zunächst unterscheiden wir, ob überhaupt Annotationen vorliegen oder nicht. Wenn Annotationen vorhanden sind, können mehrere linguistische Ebenen annotiert sein: *Morphosyntax*, *Syntax*, *Semantik*, *Pragmatik*, *Fehler*, und weitere Ebenen, auf die wir im Buch wenig oder überhaupt nicht eingegangen sind wie *Textstruktur* und *Informationsstruktur*[8], *Phonetik/Prosodie*, *Gestik* usw. Für eine genauere Erläuterung der Werte verweisen wir auf Kapitel 4.

Kriterium: Größe

Die ersten digitalen Korpora wie das *Brown Corpus*[9] umfassen ca. 1 Millionen Wortformen. Aktuelle Referenzkorpora wie das *British National Corpus* und das *American National Corpus* für das Englische und das *DWDS-Kernkorpus* für das Deutsche umfassen

---

5 Vgl. Sinclair (1996).
6 Vgl. Koch und Oesterreicher (1994).
7 Vgl. hierzu Lemnitzer und Naumann (2001), Abschnitt 4.
8 Vgl. hierzu aber die vorzügliche Einführung von Manfred Stede (2007).
9 Vgl. `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM`.

100 Millionen Wortformen. Die aktuell mögliche Größe für Korpora liegt bei mehreren Milliarden Textwörtern, wie etwa beim *Deutschen Referenzkorpus* (DeReKo) am Institut für Deutsche Sprache und einigen Webkorpora.

Die Größe eines Korpus spielt in mehrfacher Hinsicht eine Rolle. Zum einen sind bestimmte, vor allem quantitative Analysen erst möglich, wenn die zugrundeliegenden Primärdaten so umfangreich sind, dass sich in ihnen in ausreichender Zahl Daten bzw. Beispiele finden, um eine valide statistische Aussage zu gewährleisten. Ivanova et al. (2008, Abschnitt 6.2) zeigen dies am Beispiel von Wortprofilen ('word sketches'). Eine verlässliche Menge von typischen Kookkurrenzen zu einem Stichwort kommt, so zeigen die Autoren, erst zustande, wenn die Basis, für die diese Kookkurrenzen erhoben werden, mindestens 600mal im Korpus vorkommt. Je größer ein Korpus ist, desto mehr Wörter überschreiten diese Frequenzlinie und werden damit zu Kandidaten für ein verlässliches Wortprofil. In etwas allgemeinerer Weise zeigen Dan-Hee Yang und Kollegen[10] die Effekte von Korpusgrößen auf Verfahren der Extraktion lexikalischer Informationen. Ryohei Sasano und Kollegen[11] haben einige Experimente durchgeführt, in denen sie zeigen, wie sich verschiedene Korpusgrößen auf die Extraktion von Kasusrahmen (für das Japanische) auswirken.

Ein anderer Aspekt betrifft Phänomene, die so selten oder komplex sind, dass ein Korpus sehr groß sein muss, damit überhaupt ein einziges Exemplar dieses Phänomens gefunden wird. Diese Situation tritt typischerweise in korpusgestützten Untersuchungen auf, bei denen z.B. eine syntaktische Hypothese verifiziert werden soll. Das gänzliche Fehlen von Evidenz spricht nicht automatisch dafür, dass die Hypothese verworfen werden muss, wie wir in Abschnitt 3.3.2 diskutiert haben. Das Vorhandensein von Evidenz ist aber ein starker Indikator, der die Plausibilität der Hypothese stützt (mehr zur Plausibilität linguistischer Hypothesen in Abschnitt 2.2). Sehr große Korpora können hier von großem Wert sein.

Man sollte sich von der Größe eines Korpus aber nicht zu sehr bei der Auswahl eines geeigneten Korpus für die eigenen Untersuchungen beeinflussen lassen. Letztendlich hängen Design und Größe des Korpus von der gewählten Fragestellung ab. Man kann auch mit relativ kleinen Korpora interessante Untersuchungen durchführen, wie Mohsen Ghadessy et al. zeigen[12].

Kriterium: Persistenz

Die meisten Korpora sind *statische Korpora*, d.h. sie bestehen aus einer abgeschlossenen Textmenge, die in einem bestimmten Zeitraum gesammelt wurde und dann für die weitere Verarbeitung gespeichert ist. Auch *statische* Korpora müssen nicht für immer eingefroren sein. Oft arbeiten die Projekte weiter und ergänzen in bestimmten Zeitabständen das Datenmaterial. Diese Ergänzungen werden normalerweise in neuen Versionen veröffentlicht. Man muss bei Arbeiten zu statischen Korpora daher auf die Version der Korpora achten.

Der Begriff des *Monitorkorpus* stammt wahrscheinlich von John Sinclair. Er bezeichnet Korpora, deren Größe sich ändert. Der Grund für die Größenänderung kann dar-

[10] Vgl. Yang et al. (2002).
[11] Vgl. Sasano et al. (2009).
[12] Vgl. Ghadessy et al. (2001).

in liegen, dass das Korpus kontinuierlich wächst, weil zum Beispiel fortlaufend neue Ausgaben einer Tageszeitung ergänzt werden. Ein anderer Grund kann sein, dass das Korpusmaterial permanent erneuert und ausgetauscht wird, weil man aus Gründen der Effizienz und des Urheberrechts die Textdaten nur so lange speichert, bis eine Untersuchung, z.B. die Extraktion noch nicht registrierter Lexeme, abgeschlossen ist. Nachteil eines Monitorkorpus ist, dass die Ergebnisse einer Untersuchung nicht (oder nur bedingt) an dem gleichen Material wiederholt werden können.

Kriterium: Bezug zum Untersuchungsgegenstand

Was hiermit gemeint ist, erklären wir am besten anhand der Werte, die dieses Kriterium haben kann. *Referenzkorpora* sollen die Eigenschaften des dadurch repräsentierten Gegenstandes möglichst gut abdecken. Im Normalfall bedeutet *Gegenstand* hier eine natürliche Sprache in einer bestimmten zeitlichen Periode, zum Beispiel ‚das Deutsche des 20. Jahrhunderts'. Referenzkorpora dienen auch als *Kontrollkorpora* für Untersuchungen, die sich auf *Spezialkorpora* beziehen und Eigenschaften der durch dieses Spezialkorpus repräsentierten Varietät untersuchen. Die Besonderheiten der untersuchten Varietät werden sichtbar, wenn man die Verteilung der zu untersuchenden Phänomene im Spezialkorpus und im Referenzkorpus vergleicht. Auf das Verhältnis von Korpus und repräsentiertem Gegenstand bezieht sich auch das Kriterium der *Ausgewogenheit* von Korpora[13]. Ein ausgewogenes Korpus ist in sich heterogen. Das klingt zunächst nach einem Widerspruch. Es bedeutet aber nur, dass ein ausgewogenes Korpus der Heterogenität einer Sprache gerecht wird. ‚Das Deutsche' zum Beispiel existiert nicht als abgeschlossenes Ganzes. Mündliches Deutsch unterscheidet sich von schriftlichem, und bei letzterem macht es einen großen Unterschied, um welche Textsorte es sich handelt. So findet man in Gesetzestexten eine andere Sprache als in Tagebuchnotizen. Nach diesem Kriterium lassen sich auch varietätenspezifische Korpora charakterisieren, z.B. Dialektkorpora, Fachsprachenkorpora, Gruppensprachenkorpora.

Die Datenbeschaffung stellt bei wohldefinierten Designkriterien manchmal ein Problem dar. Viele Daten stehen wegen Copyright-Beschränkungen der Forschung nicht zur Verfügung und erst recht nicht für eine Veröffentlichung als allgemein zugängliches Korpus. Deshalb beruhen viele Korpora auf mehr oder weniger *opportunistischen* Datenzusammenstellungen, d.h. ein Text wurde vor allem deshalb zum Teil eines Korpus, weil er frei zur Verfügung stand. Für eine Verwendung von opportunistischen Korpora kann damit argumentiert werden, dass sowieso kein Korpus wirklich repräsentativ ist. Bei einem opportunistischen Korpus handelt es sich im Normalfall um ein Spezialkorpus, zum Beispiel mehrere Jahrgänge einer Tageszeitung. Deshalb ist opportunistisches Korpus auch nicht als eigenständiger Wert in der Korpustypologie aufgeführt.

Kriterium: Verfügbarkeit

Dieses Kriterium wird bei der Diskussion um Metadaten selten thematisiert[14], ist aber für Sie als potenzieller Nutzer von großem Interesse. Neben Korpora, die man über Online-Schnittstellen frei durchsuchen oder herunterladen kann, ist es bei kostenlosen

[13] Vgl. Atkins et al. (1992).
[14] Vgl. aber Hunston (2008), S. 157.

Korpora oft üblich, dass man sich als Nutzer registrieren oder einen (kostenlosen) Lizenzvertrag abschließen muss. Mit den Lizenzverträgen soll sichergestellt werden, dass die Daten nicht missbraucht werden, wenn sie z.B. Informationen zu Privatpersonen beinhalten. Zum anderen sollen die Daten nicht zu kommerziellen Zwecken verwendet werden, ohne dass die Ersteller des Korpus ebenfalls davon profitieren. Bei manchen Korpora sind Annotationen kostenfrei verfügbar, man muss allerdings nachweisen, dass man Lizenzgebühren für die Primärdaten bezahlt hat.

7.3 Deutsche Korpuslandschaft

Auf der Webseite, die dieses Buch begleitet, geben wir eine tabellarische Übersicht über deutschsprachige Korpussammlungen und einzelne Korpora. Diese Übersicht ist nach den genannten Kriterien strukturiert[15].

Es ist uns wichtig, hier zu erwähnen, dass wir bei der Erstauflage dieses Buches versucht hatten, eine umfassende Übersicht zu geben. Das Feld der deutschsprachigen Korpora hat sich seither erfreulicherweise sehr geweitet, so dass wir nicht mehr den Anspruch auf Vollständigkeit haben. Falls Sie Korpora kennen, die wir nicht aufführen, möchten wir Sie trotzdem bitten, uns dies mitzuteilen, damit wir das betreffende Korpus auf der Webseite ergänzen können.

In den folgenden Abschnitten werden wir noch einmal detaillierter auf die möglichen Werte der Kriterien unserer Typologie eingehen, wie Sie sie auch in der Tabelle auf der Webseite finden. Zusätzlich beschreiben wir jeweils ein paar typische Vertreter aus der deutschen Korpuslandschaft.

7.3.1 Funktionalität der Korpora

Im Zweifelsfall werden Korpora mit dem Zweck erstellt, als *empirische Basis für linguistische und/oder computerlinguistische Forschung* zu dienen. Im Bereich der Computerlinguistik gilt das oft für große, opportunistisch gesammelte Korpora, wie z.B. das *DECOW*-Korpus, das an der Freien Universität Berlin aufbereitet wurde. Fast ebenso unspezifisch ist die Angabe *Sprachdokumentation*, die für Korpora von älteren Sprachstufen verwendet wird, z.B. dem *Bonner Frühneuhochdeutsch-Korpus* dem Referenzkorpus *Mittelhochdeutsch-Korpus*, sowie das Korpus *Emigrantendeutsch in Israel* (im Deutschen Spracharchiv in Mannheim archiviert), welches eine Varietät des Deutschen der 1920er Jahre dokumentiert. Natürlich dokumentieren alle Korpora Sprache. Hier ist jedoch gemeint, dass die Korpora etwas dokumentieren, das nicht durch neue Datenerhebungen ersetzt werden kann.

Als Datengrundlage für bestimmte lexikographische Projekte wurden z.B. die *IDS Handbuchkorpora* oder das *DWDS-Kernkorpus* kompiliert. Als Material für die Sprachlehre dienen z.B. das *CG-annotierte Europarl Korpus* als auch das Lernerkorpus *Falko*. Beide Korpora wurden aber zu ganz unterschiedlichen Zwecken in diesem Bereich konzipiert. Das *Europarl*-Korpus soll als konkretes Übungsmaterial für Lerner dienen, wohingegen das *Falko*-Korpus die Sprache von Lernern dokumentiert für Untersuchungen

[15] Das Kriterium Persistenz ist in der Tabelle nicht aufgeführt. Die drei einzigen Monitorkorpora unserer Liste werden unten gesondert vorgestellt.

zu Fragestellungen des Fremdspracherwerbs und in Hinblick auf eine sprachdidaktische Auswertung. Das Dortmunder *Chat-Korpus* wurde als empirische Grundlage für Forschung im Bereich der internetbasierten Kommunikation aufgebaut. Eine konkrete computerlinguistische Motivation stand hinter der Erstellung des *Hypnotic-Korpus*. Es dient als Datengrundlage für die Programmerstellung einer automatischen Klassifizierung von Webseiten. Aus dem Bereich der gesprochenen Sprache wollen wir als Beispiel noch das *Vineta-Korpus* nennen, das für die Untersuchung von intonatorischen Verfahren zusammengestellt wurde, sowie das Dirndl-Korpus als Beispiel für die Forschung an der Schnittstelle Prosodie und Informationsstruktur.

7.3.2 Sprachenauswahl der Korpora

Die meisten der aufgeführten Korpora sind monolingual *Deutsch*. In der Datenbank des *Archivs für gesprochenes Deutsch* (AgD) am Institut für Deutsche Sprache findet man auch Dialektkorpora und Korpora von österreichischen und deutsch-schweizer Sprechern. Das C4-Korpus, das u.a. über das Digitale Wörterbuch der deutschen Sprache in Berlin verfügbar ist, setzt sich aus Teilkorpora zusammen, die hochdeutsche, schweizerische, österreichische und südtiroler Varianten des Deutschen repräsentieren.

Auch ältere Sprachstufen werden in der Tabelle explizit aufgeführt. An dieser Stelle sei auf *Mediaevum* verwiesen, ein Internetportal, das Informationen über Korpora und andere Ressourcen des Mittel- und Althochdeutschen bereitstellt. Zusätzlich wollen wir Sie auf die Referenzkorpora des Projekts *Korpus historischer Deutscher Texte* aufmerksam machen, die an verschiedenen Standorten in Deutschland zu verschiedenen Sprachstufen und Dialekten (Althochdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Mittelniederdeutsch/Niederrheinisch u.a.) erarbeitet werden.

Eine weitere Variante des Deutschen ist Lernerdeutsch, wobei hier zwischen dem Erstspracherwerb und dem Fremdspracherwerb unterschieden wird. Für Daten zum Erstspracherwerb verweisen wir auf die Datenbank *CHILDES*. Fremdspracherwerb von Erwachsenen wird z.B. im *Learning Prosody* (LeaP) Korpus dokumentiert. Im Falko-Korpus finden Sie Essays und andere Texte von Fremdsprachenlernern des Deutschen. Das LeaP-Korpus ist ein bilinguales Vergleichskorpus mit Tonaufnahmen von Nicht-Muttersprachlern in Englisch und Deutsch. Ein bilinguales Parallelkorpus liegt mit dem *INTERSECT*-Korpus aus Brighton vor, das ebenfalls das Sprachpaar Deutsch-Englisch dokumentiert, jedoch als direkte Übersetzung. Das Korpus ist auf Satzebene aligniert. Auf der *OPUS* Plattform („open source parallel corpus") findet man Korpora mit über zwanzig verschiedenen Sprachen, die auf Satzebene und zum Teil auf Wortebene aligniert sind; die Zahl der Korpora, die auf dieser Seite bereitgestellt wird, wächst kontinuierlich. Eines dieser Korpora ist das *Europarl*-Korpus, das Übersetzungen von Debatten des europäischen Parlaments in alle Amtssprachen der EU beinhaltet. In Version v3 haben Sie Zugriff auf die elf offiziellen Amtssprachen vor 2004, in Version v7 auf 21 offizielle Amtssprachen zum Stand von 2007 (alle außer Irisch). Einschränkend muss hier erwähnt werden, dass nicht alle Sprachpaare direkte Übersetzungen voneinander darstellen, sondern dass oftmals die Übersetzung von und ins Englische als Brücke genutzt wird.

7.3.3 Medium der Korpora

Unsere Sammlung legt entsprechend unserer persönlichen Forschungsausrichtungen einen Schwerpunkt auf Korpora der *geschriebenen* Sprache. Dies entspricht der Gesamttendenz unseres Buches, nicht aber dem, was an Korpusressourcen tatsächlich vorhanden ist. Für Korpora der *gesprochenen* Sprache wollen wir vor allem auf die großen Archive verweisen, das *Bayerische Archiv für Sprachsignale* (BAS) und das *Archiv für gesprochenes Deutsch* (DGD) in Mannheim, das nicht nur ein Archiv ist, sondern auch eine Möglichkeit zur Online-Recherche bietet. Sowohl geschriebene als auch gesprochene Anteile enthalten z.B. das *LIMAS*-Korpus und das *DWDS*-Kernkorpus. Im Dirndl-Korpus sind die geschriebenen Anteile keine Transkripte der vorgelesenen Radionachrichten, sondern wurden parallel mit den Audiodaten veröffentlicht. Eventuell handelt es sich dabei um die Nachrichtenmanuskripte. Zusätzlich wurde die Audiospur transkribiert Das Korpus behinhaltet somit zwei aligninerte Textversionen, eine medial schriftliche und eine medial mündliche. Die medial mündlichen Transkriptionen dokumentieren Versprecher und andere Besonderheiten der Audiospur. Das Freiburger *Videokorpus zur Aphasie* ist ein *multimodales* Korpus. Es beinhaltet Audio- und Videospuren, Transkriptionen und weiterführende Annotationen. Ähnliches gilt für das *Multilingual Soccer Corpus* von Thomas Schmidt.

Ein Sprachkorpus der besonderen Art ist das *Deutsche Gebärdensprachkorpus* (DGK), das an der Universität Hamburg entsteht. Hier dient die Videospur nicht nur zur Kontextualisierung der Sprache, sondern ist die primäre Speicherform der Gebärden, d.h. der medial visuellen Sprachzeichen.

7.3.4 Größe der Korpora

In unserer Aufstellung finden Sie einige relativ kleine Korpora, wie zum Beispiel das *Vineta*-Korpus, das Transkriptionen von nur ungefähr 46 Minuten Gespräch umfasst. Es wurde von Stefan Rabenus für seine Doktorarbeit in Einzelarbeit aufgenommen und annotiert. Das *Potsdam Commentary Corpus* (PCC) umfasst 174 Artikel einer Tageszeitung mit 32 800 Token. Hierbei handelt es sich zwar nicht um eine Einzelarbeit, die relativ geringe Größe erklärt sich aber dadurch, dass das Korpus als Pilotprojekt für die Annotation von sehr komplexen Diskursstrukturen betrachtet werden kann. Ein weiteres kleineres Korpus ist das *Learning Prosody*-Korpus (LeaP). Es ist wie das *Vineta*-Korpus ein Korpus der gesprochenen Sprache und umfasst in der Transkription ca. 76 000 Token. Es wurde ebenfalls im Rahmen eines Dissertationsprojekts erstellt.

Manche der erwähnten Korpora sind dagegen sehr groß. Das *Kernkorpus* des Projekts *Digitales Wörterbuch der Deutschen Sprache* umfasst 100 Millionen Token, zusammen mit den Erweiterungskorpora stehen hier knapp 3 Milliarden Token zur Recherche zur Verfügung. Deutlich mehr als diese 2 Milliarden Token können online beim *Institut für Deutsche Sprache* oder in Webkorpora wie dem DECOW-Korpus durchforstet werden. Beachten Sie, dass die großen Referenzkorpora der deutschen Gegenwartssprache und Webkorpora ständig erweitert werden. Es kann sich bei den hier genannten Zahlen also nur um Momentaufnahmen handeln.

7.3.5 Annotation der Korpora

Ohne jegliche linguistische Annotation kommen zum Beispiel die Rohtexte der Zeitungsverlage aus[16]. Auch in vielen Sammlungen historischer Texte findet man meistens keine weiterführende Annotationen. Die Korpora des *Instituts für Deutsche Sprache* sind mit Textstruktur annotiert. Teilweise enthalten sie auch morphosyntaktische Annotation, wie ein Großteil aller aufgeführten Korpora. Die Annotation ist in den meisten Fällen automatisch erstellt und daher mit gewissen Fehlern behaftet. Syntaktisch annotiert und manuell kontrolliert sind die beiden Korpora der Baumbankprojekte *TIGER* und *Tüba-D/Z*. Beispiele für weiterführende Annotationen, die auf der syntaktischen Annotation aufbauen, sind die semantische Annotation in der *SALSA*-Baumbank und die pragmatische Annotation im *Potsdam Commentary Corpus*. Im *Dirndl*-Korpus werden pragmatische Annotationsebenen zu Informationsstatus und Koreferenz auf der Basis einer automatischen Syntaxannotation erstellt.

7.3.6 Persistenz der Korpora

Dieses Kriterium taucht als einziges in der Übersicht nicht explizit auf. Der Grund ist, dass fast alle Korpora statische Korpora sind. Anders sieht dies bei den (z.T. online und frei verfügbaren) Archiven großer Zeitungen wie etwa der ZEIT aus. Diese wachsen kontinuierlich. Ob ältere Texte aus diesen Archiven entfernt oder hinter eine sog. Zahlschranke verbannt werden, das hängt von der Geschäftspolitik der Verlage oder anderen kommerziellen Anbieter der Archive ab.

7.3.7 Sprachbezug der Korpora

Unsere Sammlung enthält zwei *Referenzkorpora*: das *LIMAS*-Korpus und das *DWDS-Kernkorpus* des *DWDS*-Projekts. Beide sind nach sorgfältig ausgewählten Designkriterien zusammengestellt. Das *LIMAS*-Korpus orientiert sich dabei an den Kriterien, die bei der Erstellung des *Brown*-Korpus[17] verwendet wurden, so dass es 500 Textausschnitte mit je 2000 Wörtern umfasst. Das *Kernkorpus* orientiert sich hingegen eher an Kriterien, die für das *British National Corpus* entwickelt wurden. So sind hier jeweils vollständige Texte enthalten und es wurde versucht, eine balancierte Mischung verschiedener Genres und Varietäten abzudecken.

Die meisten der von uns erwähnten Korpora sind als *Spezialkorpora* zu klassifizieren, wobei sie bei ausreichender Größe durchaus auch als Referenz, z.B. für Wortlisten, eingesetzt werden können. Ein Beispiel für ein Korpus einer Individualsprache ist das Bonner *Kant-Korpus*. Exoten unter den Spezialkorpora sind zum Beispiel das *Lufthansa-Korpus* oder das *SMS*-Korpus, weil sie extrem eingeschränkte Domänen umfassen.

7.3.8 Verfügbarkeit der Korpora

In die Ressourcenübersicht auf der Webseite haben wir nur potenziell verfügbare Ressourcen aufgenommen. Manche Ressourcen, die an anderer Stelle im Buch genannt

[16] Die Rohtexte liegen ggf. als HTML-Dokumente vor und beinhalten dann Markierungen der Textstruktur.

[17] Vgl. Kučera und Francis (1967).

sind, werden Sie dort deshalb nicht finden. Das diachrone *Mainzer Zeitungskorpus* zum Beispiel, das von Carmen Scherer für ihre Promotion zum Wortbildungswandel[18] ausgewertet wurde, wird nicht erwähnt, weil es nicht digital zur Verfügung steht. Verlagskorpora, wie das *Wahrig*-Korpus[19], haben wir aus demselben Grund ebenfalls nicht aufgenommen. Korpora, die aus rechtlichen Gründen nur institutsintern genutzt werden dürfen, wie das *Leipzig/BYU Corpus of German*[20], fehlen ebenfalls. Eine Ausnahme findet sich allerdings: das *Videokorpus zur Aphasie*. Obwohl damit nur an der Universität Freiburg Forschung betrieben werden darf, haben wir es aufgenommen. Es handelt sich um ein weltweit einmaliges Korpus einer Langzeitstudie zu zehn akuten Aphasikern und ihren Familien, die über einen Zeitraum von einem Jahr nach der Entlassung des Aphasikers aus der Klinik regelmäßig auf Video aufgezeichnet wurden. Das Korpus eignet sich sehr gut für Promotionsprojekte, die allerdings aus rechtlichen Gründen an der Universität Freiburg angesiedelt sein müssen. Um den Schutz der Privatsphäre der Teilnehmer zu wahren, werden manche Daten nur in anonymisierter Form freigegeben, so z.B. das Dortmunder *Chat-Korpus*.

Verfügbarkeit kann folgende Werte haben: *frei*, wenn die Daten frei aus dem Netz kopierbar sind; *auf Anfrage*, wenn die Daten frei sind, aber nicht im Netz stehen; *online*, wenn sie kostenlos online durchbar sind (ggf. nach Registrierung); *Lizens* bedeutet im Normalfall, dass man eine Lizenzgebühr bezahlen muss. *Verkauf* schließlich bedeutet, dass man mehrere hundert Euro für die Daten zahlen muss, wie es z.B. für das *Mannheimer Korpus* des IDS der Fall ist, wenn man anstelle eines punktuellen Online-Zugriffs, die gesamte Datenmenge nutzen möchte.

7.3.9 Übersicht zu Archiven und Portalen

In Tabelle 25 haben wir eine Reihe von nationalen und internationalen Archiven und Portalen zusammengestellt, die Korpora archivieren bzw. Links auf Korpora bereitstellen. Manche der Initiativen ermöglichen den Zugriff auf die Korpora nur gegen eine Lizenzgebühr oder verlangen eine (kostenpflichtige) Mitgliedschaft[21]. Die Angaben sind alphabetisch, nicht thematisch sortiert.

---

[18] Vgl. Scherer (2005).

[19] Es wird allerdings von Universitäten im Rahmen von gemeinsamen Projekten lizenziert und kann intern genutzt werden, so z.B. an der Universität des Saarlands in Saarbrücken.

[20] Kontakt: Randall L. Jones, Brigham Young University; Erwin Tschirner, Universität Leipzig.

[21] Erkundigen Sie sich, ob Ihr Institut zum Beispiel Mitglied im *LDC* ist. Dann stünden Ihnen eine Vielzahl von Ressourcen zur Verfügung.

Name	Adresse	Kommentar
Archiv für gesprochenes Deutsch (AGD)	`http://agd.ids-mannheim.de/index.shtml`	Stellt über das Archiv gesprochenes Deutsch Korpora gesprochener Sprache zur Verfügung; in der Datenbank Gesprochenes Deutsch ist eine Online-Recherche in alignierten Transkripten möglich.
Bayerisches Archiv für Sprachsignale (BAS)	`http://www.phonetik.uni-muenchen.de/Bas/Basfomede u.html`	Archiviert Korpora gesprochener Sprache; Daten sind frei oder unter Lizenz verfügbar.
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)	`www.deutschestextarchiv.de` und `www.dwds.de`	Deutsches Textarchiv: größtes diachrones Korpus, das den Zeitraum von 1650–1900 umfasst; DWDS: ausgewogenes Referenzkorpus für die deutsche Sprache des 20. Jahrhunderts (Kernkorpus 20), dazu Erweiterungskorpora vor allem aus Zeitungsarchiven, einige Spezialkorpora. Abfrage aller Korpora über die Webseiten mit der Suchmaschine DDC.
Child Language Data Exchange System (CHILDES)	`http://childes.psy.cmu.edu/`	Internationales Archiv für Spracherwerbsdaten; Online-Suche und teilweise frei verfügbar.
CLARIN Virtual Language Observatory – Resources	`http://www.clarin.eu/content/virtual-language-observatory`	Metadaten für Sprachressourcen fast aller europäischer Sprachen. Die meisten der hier registrierten Daten sind frei abfragbar oder herunterzuladen.
Corpora List	`www.hit.uib.no/corpora/`	Internationale Mailingliste zu Korpora mit Archiv; ist ein Forum für Korpus- als auch für Computerlinguisten.
European Language Resources Association (ELRA)	`http://www.elra.info`	Internationale Organisation, die Sprachressourcen wie Korpora zur Verfügung stellt (größtenteils kostenpflichtig). Neu: ein Universal Catalogue (`http://universal.elra.info/search.php`).
Hamburger Zentrum für Sprachkorpora (HZSK)	`https://corpora.uni-hamburg.de/drupal/`	CLARIN-Zentrum, das über ein Webportal linguistische Korpora und Tools zur Verfügung stellt, u.a. die Korpora des ehemaligen SFBs 538 mit dem Thema Mehrsprachigkeit.
Institut für Deutsche Sprache (IDS)	`http://www.ids-mannheim.de/`	Größte zentrale Korpussammlung Deutschlands, über 25 Milliarden Wörter geschriebener Standardsprache (Stand 09/2014); Annotation von Textstruktur, teilweise morphosyntaktisch annotiert; Online-Abfrage mit COSMAS II.

Name	Adresse	Kommentar
Korpora des SFB 441	http://www.lingexp.uni-tuebingen.de/sfb441/corpora/index-de.html	Auflistung der Korpora, die im SFB 441 *Linguistische Datenstrukturen* erstellt wurden.
Korpora des SFB 632	www.sfb632.uni-potsdam.de/en/corpora.html	Korpora des ehemaligen SFB 632 mit dem Thema Informationsstruktur.
Korpora.org	http://www.korpora.org/	Verschiedene deutschsprachige Korpora, u.a. Texte von Kant und Frege sowie das Bonner Frühneuhochdeutsch-Korpus.
Laudatio-Repository	http://www.laudatio-repository.org/	Gut aufbereitete Metadatensammlung für tief annotierte Korpora.
Linguist List	http://linguistlist.org/	Größte internationale Mailingliste zu allen Themen der Linguistik mit großem Archiv. Natürlich werden hier auch Themen der Korpuslinguistik verhandelt.
Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu	Amerikanische Organisation (Zusammenschluss von Firmen, Universitäten und staatlichen Stellen). Manche Korpora sind nur für Mitgliedsinstitutionen erhältlich. Deutsche Korpora gesprochener Sprache sind z.B. die Katalogeinträge LDC97S43 (CALLHOME) und LDC96S51 (CALLFRIEND).
Mediaevum	http://mediaevum.de	Sehr umfangreiches Portal zu lateinischen und deutschen Texten des Mittelalters; enthält Links u.a. zu Sprachressourcen und Hilfsmitteln.
Project Gutenberg (Englisch)	http://www.gutenberg.org	Internationales Archiv mit frei verfügbaren, größtenteils englischsprachigen Büchern.
Projekt Gutenberg	http://gutenberg.spiegel.de/	Archiv mit frei verfügbaren deutschsprachigen Büchern.
Projekt Wikisource	http://de.wikisource.org/wiki/Hauptseite	Sammlung von Quellentexten, die entweder urheberrechtsfrei sind oder unter einer freien Lizenz stehen.
TalkBank	http://talkbank.org/	Internationales Archiv für Korpora gesprochener Sprache; Online-Suche und teilweise frei verfügbar.
TITUS	http://titus.uni-frankfurt.de/indexd.htm	Portal für indogermanische Text- und Sprachmaterialien an der Universität Frankfurt; bietet die Möglichkeit der Online-Recherche.

Tabelle 25: Nationale und internationale Korpusarchive, -sammlungen und Mailinglisten

7.4 Neue Korpusinitiativen

Bis vor Kurzem war der Aufbau sehr großer, ausgewogener oder opportunistischer Referenzkorpora im Zentrum der Aktivitäten der großen Institute, die sich dieser Aufgabe widmen. Man kann zumindest für das Deutsche sagen, dass es hier ein solides Fundament gibt, das in vielen Fällen für die Gewinnung linguistisch relevanter und plausibler Erkenntnisse ausreicht. Dies gilt ungeachtet dessen, dass diese Korpora ausgebaut und aktualisiert werden.

Daneben haben sich in den letzten Jahren eine Reihe von Initiativen gebildet, deren Ziel entweder der Aufbau ausreichend großer Spezialkorpora ist oder der Aufbau von (opportunistischen) Referenzkorpora, die in der Größe die bisher erstellten Korpora noch deutlich übertreffen. In Abschnitt 7.3.4 haben wir gezeigt, was mit ausreichend groß gemeint ist.

Wir stellen im Folgenden zwei Initiativen vor, die auf die Sammlung von größeren (Referenz-)Korpora für spezielle Gegenstandsbereiche abzielen (*Deutsches Textarchiv* und *Deutsches Referenzkorpus der internetbasierten Kommunikation*), sowie zwei Initiativen, die auf die Sammlung sehr großer Datenmengen abzielen (Deutsche Webkorpora und *Google Books*).

7.4.1 Das deutsche Textarchiv

Das Projekt *Deutsches Textarchiv* (DTA) baut einen Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 auf. Die Textauswahl erfolgt auf der Grundlage einer Auswahlbibliographie. Das Ziel ist, ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus zusammenzustellen.

Um den historischen Sprachstand möglichst genau abzubilden, wurden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt[22]. Die Texte wurden orthografisch normiert und linguistisch annotiert. Deshalb ist eine schreibweisentolerante und um linguistische Kategorien erweiterte Suche in den Beständen möglich[23]. Das Korpus umfasst ca. 100 Millionen Token, wird aber zur Zeit noch ausgebaut. Ein Vergleich mit ähnlich großen Referenzkorpora ist nicht sinnvoll, da a) das Deutsche Textarchiv einen viel größeren Zeitraum abdeckt als etwas das Kernkorpus des DWDS oder das British National Corpus, und b) das Korpus aus einer kleineren Anzahl langer bis sehr langer Texte zusammengestellt wurde – was Auswirkungen auf statistische Auswertungen z.B. zur Dispersion von Wörtern haben kann[24]. Das Deutsche Textarchiv besteht aus den folgenden Komponenten:

- Das Kernkorpus, dessen Aufbau in der alleinigen Verantwortung des DTA liegt.
- Eine Erweiterungskomponente[25]; das Projekt ist offen für die Aufnahme und (gemeinsame) Pflege von Texten, die Wissenschaftler im Bereich der deutschen Sprache des 16. bis 19. Jahrhunderts erfasst und digitalisiert haben. Wenn Sie dazu

[22] Alle Texte wurden unter eine „Creative-Commons"-Lizenz gestellt.

[23] Vgl. `www.deutschestextarchiv.de`.

[24] Eine Aufschlüsselung der Korpusinhalte nach Textsorten und, im Bereich der Wissenschaften, nach dem Gegenstandsbereich gibt Geyken (2013), Abschnitt 3.1.

[25] Vgl. `www.deutschestextarchiv.de/\-dtae`.

gehören sollten, dann finden Sie hier die Möglichkeit, ihre Texte einer größeren Wissenschaftlergemeinschaft zugänglich zu machen, was mittlerweile auch von vielen Institutionen, die solche Arbeiten fördern, verlangt wird. Diese Erweiterungs-Texte können auch als Spezialkorpora betrachtet und mit dem DTA-Kernkorpus als Referenzkorpus im Hintergrund auf ihre sprachlichen Spezifika hin verglichen werden. Ein Interesse auf Seiten des DTA-Projekts liegt darin, dass durch die Erweiterung des Korpuskerns mit externen Quellen das Korpus zumindest in einigen Zeitscheiben eine Größe erreicht, die die Anwendung von statistischen Verfahren interessant macht.

- Die Webseite, über die eine schreibungstolerante Suche über alle Korpusdaten möglich ist; das Rechercheergebnis, also die Konkordanz, besteht aus dem Textausschnitt, in dem der Suchausdruck vorkommt und dem zugehörigen Ausschnitt aus dem Digitalisat des Originaltextes.
- Eine Qualitätssicherungskomponente[26], d.h. eine webbasierte Anwendung, um in XML/TEI-annotierten Textdigitalisaten verschiedene Arten von Fehlern zu finden, zu kategorisieren und zu korrigieren; über diese Seite bzw. Anwendung können sich Nutzer an der Qualitätssicherung des Korpus beteiligen, indem Sie z.B. ihren Lieblingstext nach den Richtlinien des DTA korrigieren.

Das DTA steht beim Aufbau des Korpus vor den folgenden, für ein diachrones Korpus spezifischen Problemen:

- Die Texterfassung ist äußerst schwierig, da die Texte, alles Erstausgaben aus der Zeit von Anfang des 17. bis Ende des 19. Jahrhunderts, mit sehr stark divergierenden Schriftarten gedruckt wurden. Automatische Texterfassung mit *Optical Character Recognition* (OCR) kommt nicht bei allen Texten in Frage. Die Alternative ist das mehrfache Abschreiben(lassen) der Texte und der anschließende Vergleich der Ergebnisse (sog. Double Keying[27]).
- Für die strukturelle Annotation der Texte werden die *Guidelines* der Text Encoding Initiative (TEI)[28] herangezogen. Diese sind aber in vielen Fällen zu weit gefasst, um eine klare und eindeutige Abbildung zwischen dem Namen eines Elementes und dem, was mit dem Inhalt gemeint ist, zu sichern. Das DTA-eigene „Basisformat" (DTABf) beschränkt die Menge der verwendbaren TEI Elemente und Attribute und gibt, wo immer möglich, feste Wertemengen vor. DTABf ist damit eine echte Untermenge der von der TEI definierten Elemente und Attribute. Dieses TEI-konforme Basisformat wird den interessierten Wissenschaftlern zur Verfügung gestellt[29], es wird momentan von etwa zwanzig externen Projekten verwendet. Wissenschaftler, die ihre Texte in das Erweiterungsmodul einbringen wollen, sind verpflichtet, das Basisformat anzuwenden, damit volle Interoperabilität mit dem Kern des DTA gewährleistet ist[30].

---

[26] Vgl. `www.deutschestextarchiv.de/dtaq` und Geyken et al. (2012a).

[27] Haaf et al. (2013) geben einen Einblick in dieses Verfahren des Double Keyings und die Qualitätskontrolle.

[28] Vgl. `http://www.tei-c.org/`.

[29] Vgl. http://www.deutschestextarchiv.de/doku/basisformat.

[30] Geyken et al. (2012b) geben einen Einblick in Entstehung und Nutzung des Basisformats.

- Die zeitlich weit gestreuten Texte weisen einen Reichtum von Schreibungsvarianten auf. Sprachliche Einheiten, die wir intuitiv als „gleich" ansehen und auch so suchen würden, werden über die Jahrhunderte zum Teil in sehr vielen Schreibungen verwendet. Um die Suche nach solchen Einheiten im Korpus zu erleichtern, müssen deshalb die verschiedenen Schreibungen „kanonisiert" werden[31].

7.4.2 Deutsches Referenzkorpus der internetbasierten Kommunikation

Ein weiteres spezifisches Korpus, mit dem eine Lücke in der Dokumentation der deutschen Sprache geschlossen werden soll, wird derzeit im Projekt *Deutsches Referenzkorpus zur internetbasierten Kommunikation* (DeRiK)[32] aufgebaut. Am Projekt beteiligt sind Wissenschaftler der TU Dortmund, der Universität Mannheim und der Berlin-Brandenburgischen Akademie der Wissenschaften. Den Kern dieses Korpus bilden nicht beliebige Webtexte, sondern die sprachlichen Äußerungen in solchen Webgenres, die in der englischsprachigen Forschung im Forschungsfeld *Computer-Mediated Communication* (CMC) und in der deutschsprachigen Forschung unter dem Oberbegriff *Internetbasierte Kommunikation* (IBK) untersucht werden. Im Fokus stehen Kommunikationstechnologien, die auf der Infrastruktur des Internets und seiner Dienste aufsetzen und die für die Realisierung dialogischer Kommunikation konzipiert sind. Prominente Beispiele für Genres internetbasierter Kommunikation sind Chats und Instant-Messaging, Diskussions-Threads in Online-Foren und in Wikis, Threads mit Nutzerkommentaren in Weblogs, Videoplattformen (z.B. YouTube) und auf den Profilseiten sozialer Netzwerke (z.B. Facebook), die Kommunikation anhand von Twitter-Postings sowie in multimodalen Kommunikationsumgebungen.

Der Fokus von DeRiK liegt auf der schriftlichen vermittelten Sprachverwendung in der internetbasierten Kommunikation. Hierunter fallen auch Genres wie Chats, die man als konzeptuell mündlich auffasst.

DeRiK verfolgt zwei Ziele, die das Korpus von Vorläufern wie dem in unserer Korpusliste aufgenommenen Dortmunder Chatkorpus unterscheiden: a) es wird Ausgewogenheit hinsichtlich der im Internet vertretenen dialogischen Genres angestrebt. Die Referenz hierfür ist die ARD/ZDF-Online-Studie zur Internetnutzung, eine jährlich im Auftrag der beiden Fernschanstalten erhobene Nutzungsanalyse; b) es soll mit anderen Referenzkorpora wie dem Kernkorpus des DWDS kombiniert bzw. mit solchen Korpora zusammen genutzt und verglichen werden können. Deshalb wird eine strukturelle Annotation der Daten angestrebt, die sich an Standards im Allgemeinen und der Annotation des DWDS-Kernkorpus im Besonderen anlehnt[33].

Die Entwickler von DeRiK stehen beim Aufbau des Korpus vor den folgenden, für ein solches Korpus spezifischen Problemen:

- Die rechtlichen Bedingungen der Wiederverwendung von Daten wie Tweets oder Kommentaren auf Zeitungsportalen und Ratgeberseiten sind unklar. In der Commu-

---

[31] In Jurish (2013) und Jurish et al. (2014) wird das Verfahren beschrieben.

[32] Vgl. `http://www.empirikom.net/bin/view/Themen/DeRiK` und Beißwenger und Lemnitzer (2013).

[33] Weitere Details zu Aufbau und Nutzung des Korpus finden sich in Beißwenger und Lemnitzer (2013).

nity haben sich Praktiken der Erhebung der Daten für Detailstudien durchgesetzt, die rechtlich in einer Grauzone liegen. Diese Praxis ist für eine begrenzte Einzelstudie gerade noch akzeptabel, aber nicht zu empfehlen, weil die Nachnutzung der Daten und damit die Replikation von Forschungsergebnissen nicht möglich ist. Für ein Korpus, das in Gänze einer weiteren Öffentlichkeit zur Verfügung gestellt werden soll, ist dies kein gangbarer Weg. Die Entwickler von DeRiK haben sich dazu entschlossen, vom Ideal der Ausgewogenheit des Korpus entlang der Nutzungsarten und -häufigkeiten, wie sie in der Online-Studie von ARD und ZDF dokumentiert sind, abzuweichen und zunächst nur Texte zu erfassen, die rechtlich unbedenklich sind[34].

- Für die strukturelle Annotation der Texte werden die Guidelines der TEI herangezogen. Diese sind aber für die Modellierung von Dokumenten der internetbasierten Kommunikation zu eng gefasst. Zentrale Elemente von IBK-Texten wie das *Posting* können damit nicht angemessen modelliert werden. Die Projektgruppe hat deshalb eine moderate Anpassung der Guidelines vorgenommen, d.h. ein auf die Textsorte angepasstes Schema für diese Texte entworfen. Ziel ist es, diese Modifikationen zu einem Bestandteil der TEI-Guidelines zu machen[35].
- Die Anwendung von sprachtechnologischen Basiswerkzeugen wie Tokenizer und Wortarten-Annotierer ist bei Texten, die vom „Modell" eines gegenwartssprachlichen Standardtextes deutlich abweichen, weniger akkurat in ihren Ergebnissen. Gleichzeitig kommt eine händische Korrektur möglicher Analysefehler bei einem Korpus dieser Größe nicht in Frage. Stattdessen müssen die Werkzeuge an die Gegebenheiten von Texten angepasst werden, die dem gegenwartssprachlichen Standard nicht entsprechen[36].
- Schließlich sind auch die Wortarten, die das als Standard verwendete Stuttgart-Tübingen Tagset (STTS, s. Abschnitt 4.3.1) für die Annotation bereitstellt, für die Beschreibung einiger Phänome der internetbasierten Kommunikation (das sog. ‚netspeak') nicht geeignet. Thomas Barz u.a.[37] entwerfen und diskutieren notwendige Erweiterungen des Tagsets.

Bisher sind in diesem Projekt die konzeptuellen Grundlagen gelegt, es wurden Daten akquiriert und zum Teil annotiert, unter den o.g. rechtlichen Restriktionen. In 2015 sollen erste Daten aus dem DeRiK-Projekt auf der Plattform des DWDS[38] verfügbar gemacht werden.

7.4.3 Deutsche Webkorpora

Webkorpora sind sehr große, opportunistisch zusammengestellte Sammlungen von Texten (einer bestimmten Sprache), die von Webseiten heruntergeladen und so aufbereitet werden, dass die Texte für typische Benutzergruppen wie Linguisten und Lexikographen

---

34 Ausführlich hierzu Beißwenger und Lemnitzer (2013), Abschnitt 3.

35 Details über diese Arbeit und das Schema finden sich in Beißwenger et al. (2012).

36 Jurish und Würzner (2013) stellen eine solche Anpassung im Bereich der Tokenisierung und Satzsegmentierung vor.

37 Vgl. Thomas Bartz (2013).

38 www.dwds.de.

verwendbar sind. Hierzu gehören a) die Bereinigung der Texte, b) eine qualitativ ausreichende linguistische Annotation der Texte und c) eine Suchmaschine, über die die Texte abfragbar sind und in gewohnten Formaten präsentiert werden, mindestens als KWIC-Konkordanz der Treffer mit Angabe der Textherkunft (typischerweise als URL der Quelle).

Chris Biemann und Kollegen[39] nennen folgende Vorteile von Webkorpora gegenüber den bisher verfügbaren Referenzkorpora wie dem BNC und dem Kernkorpus des DWDS: a) sie sind in der Regel als Ganzes verfügbar, wohingegen auf den traditionellen Referenzkorpora rechtliche Restriktionen der Textgeber eine Weiterverbreitung des Korpus als Ganzes unmöglich machen, b) die Menge an laufenden Wörtern ist größer[40], c) die Daten sind aktueller und d) die Vielfalt der Textsorten ist größer (35–36).

In den letzten fünf Jahren sind beim Aufbau und der Bereitstellung von Webkorpora bedeutende Fortschritte erzielt worden, zugleich sind die Probleme bzw. Herausforderungen bei der Erstellung und der Nutzung dieser Art von Korpus deutlicher geworden:

- Es ist schwierig, eine ausgewogene Vielfalt von Texten aus der noch größeren verfügbaren Datenmenge des World Wide Web zu extrahieren. Die für den Zugriff auf und das Herunterladen von Webseiten verwendeten Verfahren nennen sich ‚Crawling', die verfügbaren Werkzeuge Crawler. Ein Manko der gängigen Crawlingverfahren ist es, dass sie Textsammlungen erzeugen, von denen ein Großteil von sehr wenigen Webservern stammt. Biemann et al. diskutieren die Vor- und Nachteile alternativer Crawling-Strategien (24–27).
- Die heruntergeladenen Texte müssen bereinigt werden. Einige Seiten sind ganz zu entfernen, weil sie z.B. nur Navigationselemente enthalten, andere Seiten enthalten viele Textbausteine (engl. ‚boilerplate text'), die, wenn sie im Korpus blieben, viele statistische Auswertungen verfälschten. Ein weiteres Problem ist die Zeichenkodierung. Es gibt mehrere nebeneinander verwendete Standards für die Kodierung von Texten im Web, die in einem Korpus vereinheitlicht werden müssen. Schließlich müssen Dubletten entfernt werden, also Webseiten mit identischem Text. Einige Projekte gehen so weit, Dubletten auch auf Satzebene zu entfernen (30–35).
- Die rechtliche Situation bzw. die Eigentumsverhältnisse und erlaubten Nutzungsarten vieler aus dem Web heruntergeladener Texte ist und bleibt unklar. Bei den in Webkorpora-Projekten heruntergeladenen Mengen von Texten ist es unmöglich, alle Urheber bzw. Rechteinhaber ausfindig zu machen und die Weiternutzung zu klären. Viele Projekte behelfen sich damit, dass sie die Abfolge der Sätze eines Textes durcheinanderbringen[41]. Linguistische Untersuchungen, die sich auf die (satzübergreifende) Kohärenz von Texten beziehen, sind mit solchen Korpora – oder, genauer: Satzsammlungen – nicht möglich.

---

[39] Die Darstellung in diesem Abschnitt stützt sich auf Biemann et al. (2013). Seitenzahlen in Klammern beziehen sich darauf.

[40] Wenn man den Umfang der Korpussammlungen des DWDS und am Institut für Deutsche Sprache mit dem Umfang heute verfügbarer Webkorpora vergleicht, ist dieses Argument wenig stichhaltig. Alle Sammlungen bewegen sich momentan im einstelligen oder kleinen zweistelligen Milliardenbereich an laufenden Wörtern.

[41] „To avoid legal problems with copyright claims, the published corpora are sentence shuffles.", http://hpsg.tu-berlin.de/cow/.

- Die heruntergeladenen Texte sind nicht immer Sätze der Sprache, für die man ein Korpus aufbauen möchte (Zielsprache). Die durch eine automatische Spracherkennung als nicht zur Zielsprache gehörenden Texte oder Sätze müssen entfernt werden. Auch diese Operation ist fehlerbehaftet[42].
- Es gibt zu den Texten in Webkorpora praktisch keine Metadaten, diese sind bei der Menge von Texten auch nicht mit vertretbarem Aufwand zu ermitteln oder zu rekonstruieren. Ein kleiner Test mit dem DECOW-Korpus und der Suchumgebung, wo man zu den Konkordanzzeilen zumindest die URL erhält, ergab, dass die Mehrzahl der getesteten URLs zu nicht mehr existierenden Seiten führten. Obwohl der Test keinesfalls repräsentativ war, steht zu befürchten, dass man zur Provenienz vieler Texte gar nichts erfährt. Anwender, die auch nur auf ein Minimum von Metadaten wie Entstehungsdatum oder Autor bzw. (bei Zeitungen) Quelle eines Textes angewiesen sind, können aus diesem Grund mit Webkorpora nichts anfangen.

Trotz dieser nach wie vor bestehenden Probleme sind einige der deutschen Webkorpora (die wir auch in unsere Liste der Einzelkorpora aufgenommen haben) für Linguisten, Lexikographen etc. benutzbar. Sie sind als Ganzes herunterladbar (sofern die notwendigen Rechnerkapazitäten bereitstehen) oder über Suchmaschinen abfragbar. Das deutsche COW-Korpus ist außerdem mit Wortarten annotiert und die Suchmaschine erlaubt die Suche nach Kombinationen von Wort und Wortart[43].

Sabine Schulte im Walde und Stefan Müller[44] zeigen an einem Beispiel, bei dem es um die Überprüfung und Verifizierung von Sprecherurteilen hinsichtlich semantischer Beziehungen zwischen Wörtern ging, den Nutzen von großen Webkorpora für diese Art von korpusgestützter Untersuchung[45].

7.4.4 Die Google-Familie

Mit „Google-Familie" meinen wir neben der allseits bekannten Suchmaschine eine von Google initiierte sehr umfangreiche Sammlung von gescannten und mit OCR aufbereiteten Büchern[46] und einer Anwendung, in der aus dem Bestand dieser Bücher n-Gramme

---

42 Ein kleines Experiment: Zum Stichwort *horse* findet man im Webkorpus an der FU Berlin (DECOW) 40 Treffen, die Mehrzahl ist in komplett englische Sätze eingebettet. Im deutschen Korpus des WaCky-Projekts findet man 1162 Treffer (deWaC), nach erster Durchsicht scheinen die meisten davon (Teil von) Eigennamen zu sein. Im DWDS-Kernkorpus findet man 16 Treffer, auch hier die meisten (Teil von) Eigennamen, zwei Belege entstammen einem längeren Zitat aus einem Buch namens „Amerikafahrt", womit die Bandbreite der (erwünschten und unerwünschten) Verwendungsweisen von Wörtern aus anderen als der Zielsprache gut abgedeckt ist.

43 Eine Suche nach dem Wort *sieben* mit der Wortart „finites Verb" fördert auch die (wenigen) Kontexte zutage, in denen dieses Wort als Verb verwendet wird.

44 Vgl. Schulte im Walde und Müller (2013).

45 Die verwendeten Korpora werden auf S. 87f. beschrieben. Interessant in diesem Zusammenhang ist, dass mit SdeWac ein von deWaC abgeleitetes, d.h. nachträglich bereinigtes Korpus verwendet wurde. Die Autoren schreiben außerdem, dass vor allem die Korpusgröße einen Einfluss auf die Qualität der Untersuchungsergebnisse hatte (S. 100).

46 S. `http://books.google.de`.

bis zu einer Länge von fünf Wörtern und deren Vorkommenshäufigkeiten im Verlauf der Zeit, aus der die Bücher stammen, abgefragt werden können[47].

Besonders die letzte Anwendung lässt die sprachtechnologischen Ambitionen von Google deutlich werden und gibt zu der Hoffnung Anlass, dass von hier aus Ressouren und Werkzeuge entstehen, die für korpuslinguistische Forschung interessant sind[48].

Während wir in den letzten drei Abschnitten dargestellt haben, welche Herausforderungen diejenigen zu bewältigen haben, die die jeweilige Ressource aufbauen wollen, müssen wir bei den Ressourcen der Google-Familie darstellen, welche Herausforderungen diejenigen zu bewältigen haben, die diese Ressourcen (für korpuslinguistische Studien) nutzen wollen. Adam Kilgarriff spricht von „Googleology" als schlechte Art, Wissenschaft zu betreiben[49]: Da die Ressourcen und Werkzeuge nicht primär für die Nutzung durch Korpuslinguisten entwickelt wurden, liegt es in der Verantwortung dieser Nutzer, zu gewährleisten, dass die auf Grund dieser Ressourcen gewonnenen Ergebnisse wissenschaftlichen Grundanforderungen wie Verlässlichkeit und Plausibilität genügen und vor allem replizierbar sind. Kilgarriff stellt fest – mit Gültigkeit für die Zeit vor dem Jahr 2007 und beschränkt auf die Google-Suchmaschine –, dass man neben der eigenen Wissenschaft auch noch Googleologie betreiben müsste, um diese Ziele zu erreichen. Kilgarriffs kritischer Einwurf kann auch als einer der Anstöße für die Korpuslinguistik gesehen werden, die Sache selber in die Hand zu nehmen und selber große Webkorpora aufzubauen und eine den Anforderungen von Korpuslinguisten genügende Infrastruktur um diese Ressourcen herum bereitzustellen (s. voriger Abschnitt). Mark Davies[50] argumentiert auf einer ähnlichen Linie wie Adam Kilgarriff, wenn er zeigt, welche Art von Fragen (die man als Korpuslinguist gerne stellen möchte) man an ein Referenzkorpus wie das *Corpus of Contemporay American English*, nicht aber über die Google-Suchmaschine stellen kann. In noch deutlicherer Weise und an Hand vieler Beispiele demonstriert Dominik Brückner[51] am Beispiel der Benutzung von Google Books für den Zweck, Belege für ein diachron ausgerichtetes Wörterbuch zu finden, dass diese Ressource und die Infrastruktur hierfür ungeeignet sind, da a) die Ergebnisliste(n) willkürlich sind und sich je nach Abfrageart und in nicht nicht nachvollziehbarer Weise ändern; b) die Metadaten auch durch die schlechten Resultate der automatischen Textrerfassung nicht den minimalen dokumentarischen Standards genügen und c) in der Dokumentation (so vorhanden) selbst grundlegende Dinge wie die, was unter einem Buch oder einem Suchergebnis verstanden wird, nicht erläutert sind.

Dem stehen andere Ansichten und Erfahrungen gegenüber, die nicht verschwiegen werden sollen. Stefan Diemer[52] zeigt an einer Fallstudie, die das Aufkommen einer neuen Art von präfigierten Verben in der englischsprachigen internetbasierten Kommunikation betrifft, dass eine ausreichende Menge von Daten nur über die Google-Suchmaschine, nicht aber über Referenz- und Webkorpora zu erhalten war. Es zeigt sich zumindest an dieser Studie, dass die Daten und Werkzeuge der Google-Familie vor allem

---

47 Vgl. `https://books.google.com/ngrams/`.
48 Vgl. Lin et al. (2012).
49 Kilgarriff (2007).
50 Vgl. Davies (2011).
51 Vgl. Brückner (2012).
52 Vgl. Diemer (2011).

dort einen guten Dienst leisten, wo die zu untersuchenden Phänomene zu selten (oder zu neu) sind, als dass sie von Referenz- und Webkorpora schon erfasst werden könnten. Auch in der empirisch orientierten Kulturwissenschaft scheint vor allem der Ngram-Viewer mit seiner Möglichkeit, sprachliche Daten für historische Längsschnittstudien zu liefern, großer Beliebtheit zu erfreuen. Diese Art der Forschung und ihre Würdigung liegen außerhalb des Rahmens dieser Einführung, deswegen wollen wir hier nur auf die Arbeit von Philipp Sarasin[53] hinweisen, in der sich auch viele anschauliche Beispiele für die Verwendung des Ngram-Viewers befinden.

Zusammenfassend kann man sagen, dass die Ressourcen und Werkzeuge der Google-Familie korpuslinguistische Nutzungen nicht ausschließen, aber auch nicht unbedingt befördern. Die Last, bei Nutzung dieser Ressourcen wissenschaftliche Standards zu wahren, liegt beim Benutzer, und diese Last wiegt hier schwerer als bei den anderen hier beschriebenen, genuin linguistischen Korpora. Google geizt mit Informationen zu seinen Ressourcen und Werkzeugen und zu allem, was direkt die Firmenpolitik berührt. In manch glücklichem Fall mögen die Interessen der linguistischen Forschungsgemeinschaft und die Interessen von Google in die gleiche Richtung gehen, das wird man aber eher als Koinzidenz zu werten haben. In vielen Fällen bedeutet die explorative Verwendung dieser Ressourcen und Werkzeuge, dass man deren Nutzung, wenn man wissenschaftliche Standards wahren will oder muss, am Ende aufgibt und sich nach geeigneteren Ressourcen umsieht. Auch dies kann eine interessante Erfahrung sein.

7.5 Weiterführende Literatur

Wir wollen an dieser Stelle nicht auf weitere Literaturstellen verweisen, sondern auf Mailinglisten, bei denen Sie sich anmelden (*subskribieren*) können. Als angemeldeter Benutzer erhalten Sie alle an diese Listen gesendeter Beiträge. Durch diese Beiträge können Sie sich über Entwicklungen in der Korpuslinguistik auf dem Laufenden halten. Sie können sich auch selbst beteiligen und dort Fragen stellen. Wenn Sie freundlich fragen, werden Sie in den meisten Fällen auch freundliche Antworten erhalten. Die Mailinglisten sind *Corpora* (`http://clu.uni.no/corpora/`, folgen Sie dem Link zur *Info Page*, dort erfahren Sie, wie Sie Mitglied werden können), *Gesprächsforschung* (`http://www.gespraechsforschung.de/liste.htm`, mit online-Registrierung) und die *Linguist List* (`http://www.linguistlist.org/`).

[53] Vgl. Sarasin (2012).

8 Wie man in den Wald hineinruft ... — Korpuslinguistik in der Praxis

Nach Lektüre dieses Kapitels werden Sie in der Lage sein, selbstständig korpusbasierte linguistische Untersuchungen entsprechend den genannten Vorbildern zu planen und durchzuführen.

8.1 Übersicht

In diesem Kapitel wollen wir linguistische Untersuchungen vorstellen, die auf deutschsprachigen Korpora basieren. Wir glauben, dass man aus diesen Beispielen etwas lernen kann – im positiven wie im negativen Sinn. Ein Blick in die germanistischen Fachzeitschriften zeigt, dass in den letzten Jahren erstaunlich viele linguistische Arbeiten entstanden sind, denen Korpora zugrunde liegen. Diese Arbeiten sind freilich von recht unterschiedlicher Qualität, wie wir noch sehen werden. Sie sind auch thematisch weit gestreut.

Wir stellen hier Projekte und Untersuchungen vor, in denen Korpusdaten eine Schlüsselrolle bei der Bearbeitung der Untersuchungsfragen spielten. Dabei berücksichtigen wir sowohl korpusbasierte als auch korpusgestützte Arbeiten, s. Abschnitt 2.4. Bei der Zuordnung der einzelnen Untersuchungen orientieren wir uns zunächst an den linguistischen Beschreibungsebenen und wählen einige für Korpusarbeiten interessante aus: Orthographie, Morphologie und Wortbildung, Syntax (Abschnitte 2–4). Dem folgen Arbeiten aus einigen Feldern der Linguistik, bei denen Korpora als Quelle der Evidenz schon immer eine besondere Rolle gespielt haben: Lexikographie und Lexikologie, Computerlinguistik und Fremdsprachenerwerb und -vermittlung (Abschnitte 5–7). Eine kritische Würdigung von neueren korpuslinguistischen Arbeiten rundet dieses Kapitel ab.

8.2 Orthographie

Der deutschen Rechtschreibung liegt spätestens seit den Zeiten Konrad Dudens eine Norm zugrunde. Deshalb der Ausdruck *Rechtschreibung*. Diese Norm, die allerdings auch Wandlungen und Reformen unterliegt, wird in der Schule vermittelt, sie kann nicht verhandelt werden. Sie ist allerdings öfter der Gegenstand von Diskussionen, was gerade die meist staatlich verordneten und durchgeführten Rechtschreibreformen zeigen. Im fachlichen Diskurs der Linguistik stehen die Prinzipien hinter der Norm sowie

Fragen ihrer Angemessenheit, Schlüssigkeit und Lernbarkeit zur Diskussion. Linguisten nehmen aktiv Anteil an der Weiterentwicklung dieser Norm, dabei werden sie von Zeit zu Zeit von interessierten Laien begleitet oder auch bekämpft.

Dementsprechend befasst sich das Gros der Arbeiten zur Orthographie-Norm mit den folgenden Themen:

- Darstellung, Begründung oder Kritik der Norm;
- Präsentation der Norm, als Menge von Regeln und / oder als Liste von Einzelwörtern;
- Vermittlung der Norm im Fremdsprachunterricht[1].

Diese Themen beziehen sich auf die festgesetzte Norm und nicht auf den tatsächlichen Sprachgebrauch. Empirische Untersuchungen an authentischen Sprachdaten sind hier überflüssig. Die Liberalisierung der orthographischen Norm im Zuge der letzten Rechtschreibreform 2006 macht es nun allerdings interessanter, am tatsächlichen Sprachgebrauch zu untersuchen, welche (zulässigen, aber auch nicht-zulässigen) Varianten in welchen quantitativen Verhältnissen verwendet werden. Darüber hinaus gibt es bei einigen Textsorten orthographische Besonderheiten, die nicht Teil der Norm sind. Beide Aspekte der orthographischen Praxis sind Gegenstand jüngerer empirischer Untersuchungen.

Günter Starke[2] vergleicht die Verwendung des Bindestrichs, vor allem des Erläuterungs- und des Durchkopplungsbindestrichs[3], mit den orthographischen Regeln und Einträgen des Rechtschreibdudens. Er stellt eine deutliche „Kluft zwischen Usus und kodifizierter Norm" fest[4]. Das Korpus, auf das er seine Untersuchungen stützt, besteht aus vier Ausgaben der Zeitschrift *Spiegel* von 1992, die der Autor, so ist zu vermuten, manuell ausgewertet hat. Die Beispiele aus diesem Korpus werden ergänzt durch Beispiele aus dem Rechtschreibduden und aus Monographien und Aufsätzen zu diesem Thema. Die Beispiele dienen dem Autor vor allem dazu, die Bereiche zu veranschaulichen, in denen Norm und Sprachgebrauch sich auseinander entwickeln. Der Aufsatz Starkes ist ein frühes Beispiel für empirische Arbeiten, in denen eine Sprachnorm systematisch mit dem Sprachgebrauch kontrastiert wird. Charakteristisch für viele empirische Arbeiten dieser Art ist die korpuslinguistische Methode: Eine kleine Textsammlung wird manuell ausgewertet. Das Material kann daher nur mehr exemplarisch sein. Die Reproduktion dieser Studie und die Erstellung einer Vergleichsstudie sind praktisch nicht möglich.

Helmut Langner[5] untersucht den Wortschatz der Sachgruppe Internet auf morphologische, aber auch orthographische Besonderheiten. Er stellt fest, dass bei der Schreibung von Wörtern aus diesem Bereich orthographische Unsicherheiten deutlich werden: „Erstaunlich ist ... das starke Schwanken zwischen Zusammenschreibung und Schreibung

---

1 Vgl. z.B. die Sammelbände von Augst (1997) sowie Eroms und Munske (1997).

2 Vgl. Starke (1993).

3 Der Erläuterungsbindestrich steht in Komposita hinter Initialwörtern und Zahlen (*BVB Desaster, 4-törig*). Der Durchkoppelungsbindestrich verbindet die Teile einer Wortgruppe zu einem Kompositum-Erstglied (*Hals-Nasen-Ohren-Arzt*), vgl. Poethe (2000).

4 Starke (1993), S. 51. Die kodifizierte Norm ist bei der Abfassung des Artikels die der zwanzigsten Auflage des Rechtschreibdudens von 1991, also noch vor der Rechtschreibreform. Die Reform hat die Norm tatsächlich etwas stärker dem *Usus* angepasst.

5 Vgl. Langner (2001). Seitenzahlenangaben in Klammern beziehen sich auf diesen Text.

mit Bindestrich, nicht selten sogar im selben Text ... Probleme haben Schreiber offensichtlich dann, wenn die Lexeme Konstituenten besitzen, die noch als fremdsprachig empfunden werden." (105) Langner stützt seine Beobachtungen auf eine Belegsammlung, die er im Jahr 2000 aus verschiedenen Quellen, vor allem Zeitung und Rundfunk, zusammengestellt hat (97). Die Beobachtungen Langners zeigen, dass sich nicht alles in einer Rechtschreibnorm regeln lässt und manche Konzepte, wie das der Fremdworthaftigkeit mancher Ausdrücke, unscharf sind. Die reformierte Rechtschreibnorm trägt dem durch eine höhere Zahl an zugelassenen Varianten Rechnung. Dennoch wird es immer orthographische Probleme jenseits der Norm geben.

Christa Dürscheid untersucht zwei Typen von „Schreibungen, die in der Rechtschreibnormierung nicht geregelt sind"[6]. Es handelt sich dabei um die Binnengroßschreibung (z.B. *InterCity*) und um die Getrenntschreibung von Komposita (z.B. *Progamm Entwickler*). Ihre These lautet, dass sich in diesen Bereichen in der Sprachverwendung Tendenzen zeigen, die früher oder später die Rechtschreibnorm verändern werden. Sie stützt ihre Analysen auf unsystematisch gesammelte Belege aus verschiedenen Medien: Fernsehen, Radio, Zeitung, aber auch aus der Beschreibung von Software oder aus der Bahnwerbung.

In einer anderen Arbeit[7] untersucht Dürscheid Verstöße gegen die orthographische Norm an verschiedenen Textsorten, die Bestandteil computervermittelter Kommunikation sind. Die Daten, auf die sie diese Untersuchungen stützt, sind Mitschnitte von Chats sowie E-Mails. Ob die nicht-normgerechten Schreibweisen in der internetbasierten Kommunikation, die nicht auf technisches oder menschliches Versagen zurückzuführen sind, Auswirkungen auf die Schreibnorm und die Schreibpraxis außerhalb dieses Mediums haben werden, kann nicht vorausgesagt werden. Die Autorin fordert hierzu weitergehende empirische Untersuchungen. Dem kann man sich nur anschließen. Es wäre wünschenswert, wenn sich solche Untersuchungen auf ein öffentlich zugängliches Referenzkorpus der computervermittelten Kommunikation stützen könnten. Ein solches ist an der Universität Dortmund und an der Berlin-Brandenburgischen Akademie der Wissenschaften im Aufbau[8].

In verschiedenen Arbeiten, die um die Jahrtausendwende herum entstanden sind[9], werden vor allem graphostilistische Elemente in internetbasierten Kommunikation, und hier vor allem bei E-Mail und Chat, untersucht: Smileys, Sonderzeichen wie Stern (*) und at Zeichen (@), prononcierte Großschreibung ganzer Wörter. Der Bereich ist für diese Formen der Kommunikation recht gut untersucht und auch solide korpuslinguistisch fundiert. Es wird in Zukunft zu zeigen sein, ob sich auch in Texten anderer neuer Medien, wie den über Mobiltelefone verbreiteten SMS, orthographische Sonderformen etablieren. SMS-Texte dürften allerdings wesentlich schwieriger zu akquirieren sein als Texte, die über das World Wide Web verbreitet werden[10]. Es gibt dennoch einige korpusbasierte Arbeiten zu diesem Thema, z.B. Schwitalla (2002), Doering (2002), allerdings

---

6 Vgl. Dürscheid (2000b), S. 223.
7 Vgl. Dürscheid (2000a).
8 Vgl. Kapitel 7.
9 Vgl. Haase et al. (1997), Runkehl et al. (1998), Storrer (2000), Storrer (2001).
10 Eine von der Universität Louvain in Belgien ausgehende Initiative baut in mehreren Ländern, u.a. in der Schweiz, zurzeit größere SMS-Korpora auf, den aktuellen Stand der Arbeiten erfahren Sie unter `http://www.sms4science.org/`.

beziehen sich diese Arbeiten nicht auf die Themen Rechtschreibnorm und Rechtschreibpraxis.

In dem Maße, wie größere Korpora mit nicht-standardsprachlichen Texten entstehen – zu nennen sind hier neben Texten der internetbasierten Kommunikation auch diachrone Korpora – stellt sich auch die Frage des Zusammenhangs zwischen standardkonformen Schreibungen und davon abweichenden Schreibungen. Im Fokus steht die graphische Einheit und Diversität des sprachlichen Zeichens. Einerseits kann das Forschungsinteresse auf den abweichend geschriebenen Formen liegen und diese Formen sollen in Korpora gefunden werden, z.B. wenn Besonderheiten der internetbasierten Kommunikation analysiert werden. Andererseits kann das Forschungsinteresse auf bestimmten sprachlichen Zeichen liegen, und das unabhängig von deren konkreter Schreibung, z.B. wenn der Wandel im Sprachgebrauch sprachlicher Zeichen über einen längeren Zeitraum untersucht werden sollen. Um beiden Suchinteressen – der Suche nach sprachlichen Zeichen in einer bestimmten grafischen Form und der Suche nach sprachlichen Zeichen unabhängig von deren grafischer Form – gerecht zu werden, wird mittlerweile bei der Annotation von Korpora mit einem hohen Anteil nicht-standardkonformer Schreibungen der ursprünglichen Form in den Primärdaten eine normaliserte Schreibung zur Seite gestellt. Diese Annotation normalisierter Wortformen bildet einen weiteren Index, über den die Korpora durchsucht werden können. Zu nennen sind hier im Zusammenhang mit diachronen Korpora die Arbeit von Bryan Jurish, Christian Thomas und Frank Wiegand[11] und im Zusammenhang mit SMS-Korpora die Arbeit von Simone Ueberwasser[12]. Die von Ueberwasser vorgestellte Annotation von überwiegend dialektalen Ausdrücken mit ihrem hochsprachlichen Pendant ist manuell vorgenommen worden. In der Arbeit von Jurish et al. wird hingegen ein Verfahren der automatischen Annotation dargestellt. Dieser Ansatz ist sicher der interessantere, da er sich auch auf große Korpora anwenden lässt. Andererseits kann dieser Ansatz nicht vollkommen fehlerfrei sein, und eine hohe Genauigkeit der Abbildung kann nur im Dialog zwischen Computerlinguisten und Philologen erzielt werden. Auf diesem Gebiet ist noch viel Spielraum für weitere Forschungs- und Entwicklungsarbeiten, da auch mehr diachrone Korpora der Forschergemeinschaft verfügbar gemacht werden.

8.3 Wortbildung

8.3.1 Aspekte der Wortbildung

Die Wortbildung ist der kreativste Bereich einer Sprache. Sprecher schaffen auf diese Weise unzählige neue Wörter, von denen viele nur dem einen, momentanen kommunikativen Zweck dienen und danach nie wieder verwendet werden (sog. Gelegenheitsbildungen).

Die Bausteine, aus denen im Deutschen neue Wörter geformt werden, sind:

- Wortstämme (z.B. *seh*, *Mutter*); eine Unterklasse der Stämme, die nicht selbständig ein Wort bilden können, wird *Konfix* genannt (z.B. *schwieger*, *thek*).

[11] Vgl. Jurish et al. (2014).
[12] Vgl. Ueberwasser (2013).

- Affixe, die nach ihrer Stellung zum Wortstamm unterschieden werden in Präfixe (z.B. *be-*), Suffixe (z.B. *-bar*) und Infixe (z.B. das Fugenelement *-s-*).
- Zwischen diesen beiden Klassen stehen Elemente, die sich von selbständigen Wortstämmen zu Affixen entwickeln, unter Verlust eines eigenen semantischen Gehalts (z.B. *-mäßig* in Wörtern wie *gefühlsmäßig*). Diese Bausteine werden in der neueren Literatur *Affixoide* genannt.
- Flexive, die grammatische Merkmale eines Worts wie Kasus oder Tempus markieren (z.B. *-en*, das als Flexiv die Infinitivform und die erste und dritte Person Plural eines Verbs markieren kann).

Ziel der Wortbildungsforschung als linguistischer Disziplin ist es, die Regeln und Beschränkungen zu formulieren, denen die freie Kombination dieser Bausteine unterliegt, und die Merkmale der aus der Kombination der Bausteine entstehenden Wortbildungsprodukte zu beschreiben. Zum Beispiel

- darf das Suffix *-bar* nur mit verbalen Wortstämmen kombiniert werden. Das entstehende Wort wird als Adjektiv verwendet. Der Beitrag des Suffixes zur Gesamtbedeutung des Adjektivs ist meist, dass die durch den verbalen Stamm beschriebene Handlung dem Gegenstand, auf den sich das neue Adjektiv bezieht, als Potenzial zugeschrieben wird (X ist ableitbar → X kann abgeleitet werden);
- kann in manchen Fällen zwischen die zwei Bestandteile eines Kompositums ein Fugenelement treten. Die Notwendigkeit des Fugenelements wird phonologisch begründet, es macht den Übergang vom letzten Phonem des ersten Wortstamms zum ersten Phonem des zweiten Wortstamms leichter (z.B. *Arbeit-s-amt*, *Tag-e-bau*). Zur Entwicklung des Gebrauchs von Fugenelementen für neue Wörter im Verlauf des 20. Jahrhunderts haben Damaris Nübling und Renata Szczepaniak eine interessante, korpusbasierte Studie vorgelegt[13].

Die Wortbildung als produktiver Prozess des Sprachausbaus steht im Spannungsverhältnis zum Lexikon einer Sprache. Wenn täglich Dutzende von neuen Wörtern gebildet werden, dann kann das Lexikon einer Sprache oder eines einzelnen Sprechers niemals vollständig in Hinblick auf das Vokabular der Sprache sein. Es ist deshalb ähnlich wie in der Syntax eine wichtige linguistische Aufgabe, die Regeln zu beschreiben, denen dieser kreative Prozess unterliegt[14]. Diese Regeln steuern die Produktion neuer Wörter und ermöglichen es den Hörern, neue Wörter korrekt zu interpretieren[15].

Empirische Sprachdaten sind auch für Wortbildungsforschung wichtig:

---

[13] Vgl. Nübling und Szczepaniak (2011).

[14] Dass in diesem Teil der Sprache Regeln wirken, sieht man an Bildungen wie *unkaputtbar*, die deshalb so auffällig sind, weil sie gegen diese Regeln verstoßen. In diesem Beispiel ist das Ziel des Regelverstoßes, Aufmerksamkeit zu erregen, und dies ist sicher gelungen.

[15] Oftmals ist dafür aber auch ein größerer Kotext oder Kontext erforderlich, wie das Beispiel *BVB-Transfer* zeigt. Ob ein BVB transferiert wird oder ein BVB etwas transferiert, erschließt sich, wenn man weiß oder erfährt, dass der BVB ein Fußballverein ist, der seine Mannschaft durch Transfers von Spielern verändert.

- Große Korpora enthalten viele Belege für die meisten Wortbildungsmuster und durchweg mehr Beispiele, als ein Wörterbuch verzeichnen kann. Gerade die nicht in Wörterbüchern verzeichneten, kontextuell gesteuerten Gelegenheitsbildungen sind ein guter Prüfstein für theoretische Annahmen zu Regeln, Regularitäten und Beschränkungen in der Wortbildung.
- Viele Wortbildungsprodukte werden erst verständlich und interpretierbar, wenn man den Kotext sieht, in dem das Wort verwendet wird. Besonders Komposita bedürfen oft der Stützung durch den Kotext[16].

8.3.2 Qualitative Untersuchungen

In den letzten Jahren ist eine Reihe von korpusbasierten Fallstudien zu einzelnen Wortbausteinen erschienen. Hierzu gehören Arbeiten von Angelika Feine sowie von Anke Lüdeling und Stefan Evert zur nicht-medizinischen Verwendung von *-itis*-Kombinationen[17], eine Arbeit von Nikolaus Ruge zum Suffixoid *-technisch*[18], eine Studie zur Valenz der *be*-präfigierten Verben von Piklu Gupta[19] sowie ein Aufsatz von Annette Klosa zu Verben mit dem Präfix *gegen-*[20].

Wir wollen in diesem Abschnitt exemplarisch die Arbeit von Susanne Riehemann zur Beschreibung der Adjektive mit dem Suffix *-bar* vorstellen[21]. Riehemann versucht anhand von intensiven Korpusrecherchen die Wortbildungsregeln und -beschränkungen im Zusammenhang mit der Verwendung des Suffixes *-bar* zu erfassen und in der Lexikonkomponente des Grammatikformalismus *Head-Driven Phrase Structure Grammar* (HPSG) zu beschreiben (2–3). Ihre Arbeit ist damit sowohl für die theoretische Linguistik als auch für die Computerlinguistik von Interesse.

Riehemann stützt ihre Untersuchungen auf neun Korpora, ein großes und acht kleinere, mit insgesamt knapp 18 Millionen laufenden Wörtern (Token). Die Frequenzangaben zu den *-bar*-Adjektiven bezieht die Autorin ausdrücklich nur auf das mit 10,7 Millionen Token größte Korpus, das Zeitungskorpus des Instituts für deutsche Sprache in Mannheim. Die kleineren Korpora bezeichnet sie als zu wenig repräsentativ, um quantitative Aussagen darauf zu stützen (5). Im einzelnen untersucht sie die folgenden Aspekte:

- Die Klassen von *-bar*-Ableitungen, vor allem hinsichtlich der zugrunde liegenden Verben. Riehemann berücksichtigt die Frequenzverteilung dieser Adjektive, die das typische Profil aller produktiven sprachlichen Prozesse aufweist: Es gibt wenige hochfrequente Wörter, die weit über die Hälfte aller vorkommenden Wörter ausmachen, und sehr viele selten vorkommende Wörter (9–12);

---

16 Auf den Zusammenhang hat Corinna Peschel in ihrer Monographie zum Verhältnis von Wortbildung und Textkonstitution hingewiesen, vgl. Peschel (2002).

17 *Handyritis, Aufschieberitis* etc., vgl. Feine (2003) und Lüdeling und Evert (2004). Auf die Arbeiten von Lüdeling und Evert werden wir im nächsten Abschnitt genauer eingehen.

18 Vgl. Ruge (2004), interessant sind hier weniger die transparenten Bildungen wie *verfahrenstechnisch*, sondern vielmehr neudeutsche Bildungen wie *gefühlstechnisch*.

19 Vgl. Gupta (2000).

20 Vgl. Klosa (2003), die Untersuchungen basieren auf den Korpora des Instituts für deutsche Sprache und auf dem DWDS-Korpus.

21 Vgl. Riehemann (1993). Die Seitenzahlen in Klammern verweisen auf diesen Text.

- die Form und Funktion der Wortbildungsprodukte, also der so entstandenen Adjektive, wobei sie vor allem deren syntaktische (mögliche Komplemente der Adjektive) und semantische Eigenschaften betrachtet (5–9);
- in einem weiteren Abschnitt diskutiert Riehemann syntaktische, semantische und pragmatische Beschränkungen des Wortbildungsprozesses, die erklären, warum einige Bildungen ungrammatisch sind, wohingegen andere, ebenfalls vom prototypischen Muster – mit einem transitiven Verb als Basis – abweichende Wörter durchaus bildbar sind (z.B. *abbaubar* mit einem intransitiven Verb als Basis und *verformbar* mit einem reflexiven Verb als Basis.) (12–16);
- Riehemann zieht auch die Argumente der zugrunde liegenden Verben in Betracht, die von dem abgeleiteten Adjektiv „ererbt" werden (*Ein Auto nach Deutschland importieren → Ein nach Deutschland importierbares Auto*). Vor allem bei der Bestimmung von Beschränkungen hinsichtlich der Vererbung von Argumenten erweist sich der Blick in das Korpus als sehr hilfreich (17–19);
- schließlich beschreibt Riehemann Unterschiede im attributiven und prädikativen Gebrauch dieser Adjektive.

Im zweiten, dem Hauptteil der Arbeit entwickelt Riehemann eine formale Beschreibung der lexikalischen Eigenschaften dieser Adjektivgruppe im Rahmen eines HPSG-Lexikons, die all den im ersten Teil der Arbeit beschriebenen Generalisierungen gerecht wird. Die Arbeit endet mit zwei Anhängen, in denen zum einen alle im Korpus vorgefundenen *-bar*-Adjektive, zum anderen die häufigsten 300 Adjektive in der Reihenfolge ihrer Häufigkeit aufgelistet sind (70–78). Riehemanns Arbeit ist ein wichtiger Beitrag zu einer formalen Beschreibung von Wortbildungsprozessen am Beispiel des vermutlich produktivsten Suffixes der deutschen Sprache.

8.3.3 Qualitativ-quantitative Untersuchungen

In jüngster Zeit ist in verstärktem Maße die Produktivität von Wortbildungselementen, wie z.B. dem Suffix *-bar*, untersucht worden. Die Produktivität in der Wortbildung hat einen qualitativen und einen quantitativen Aspekt. Beide erfordern unterschiedliche Analysemethoden.

- Der qualitative Aspekt hängt zusammen mit der Menge der Elemente, mit denen ein bestimmtes Morphem kombiniert werden kann. So ist z.B. der Anwendungsbereich des Suffixes *-bar* auf verbale Basen beschränkt, und hier fast ausschließlich auf die transitiven Verben. Das Suffix *-sam* hingegen tritt zusammen mit verbalen Basen (*arbeit-sam*) und mit adjektivischen Basen (*selt-sam*) auf. Der Anwendungsbereich von *-bar* und damit die Menge der hiermit bildbaren Wörter ist also beschränkter als der Anwendungsbereich von *-sam*;
- der quantitative Aspekt der Wortbildung kann informell beschrieben werden als die Wahrscheinlichkeit, mit der man einem mit einem bestimmten Morphem gebildeten neuen Wort begegnet, nachdem man bereits eine bestimmte Anzahl von Wörtern beobachtet hat. In einer anderen Sichtweise wird der Produktivitätsindex bestimmt von der relativen Anzahl der Wörter, die bisher nur einmal in den beobachteten

Daten auftauchten[22]. In dieser Interpretation wird man nach Analyse eines Korpus der deutschen Gegenwartssprache feststellen, dass das Suffix *-bar* relativ produktiv ist, die Produktivität des Suffixes *-sam* hingegen gegen null tendiert. Mit anderen Worten, die Wörter mit dem Suffix *sam* sind vollständig aufzählbar.

Wie man an den obigen Beispielen sieht, sind der qualitative und der quantitative Aspekt der Produktivität von Wortbildungselementen unabhängig voneinander. Die qualitative Analyse kann anhand einer Belegsammlung durchgeführt werden. Für die quantitative Analyse ist die Analyse eines kompletten, möglichst großen Korpus allerdings zwingend notwendig. Dies hat zwei Gründe:

- Erstens kann man im Hinblick auf Vorkommenshäufigkeiten von Wörtern oder Wortbildungsmustern weder die eigene Intuition noch die Intuition anderer Muttersprachler zu Rate ziehen. Hinsichtlich quantitativer Verhältnisse ist unser Sprachgefühl zu unzuverlässig;
- zweitens muss man für die hier zur Diskussion stehende Analyse eine große Menge von Texten sukzessive nach der Anzahl und Häufigkeit der Vorkommen eines bestimmten Musters durchforsten.

Anke Lüdeling und Stefan Evert[23] untersuchen den quantitativen Aspekt der Produktivität des Suffixes *-lich*. Sie verwenden hierfür ein Zeitungskorpus von ca. 3 Millionen laufenden Wörtern. Die Analyse der Klasse aller mit *-lich* gebildeten Wörter ergibt ein ziemlich unscharfes Bild. Die Analyse wird aber präziser, nachdem die Autoren vier verschiedene Klassen gebildet haben: a) *-lich* mit adjektivischer Basis (z.B. *grün-lich*), b) *-lich* mit verbaler Basis (z.B. *vergess-lich*), c) *-lich* mit nominaler Basis (z.B. *arzt-lich*) und d) *-lich* mit phrasaler Basis (z.B. *vorweihnacht-lich*). Die Kombination des Suffixes mit nominaler Basis ist sehr produktiv, die Kombination mit verbaler Basis hingegen unproduktiv. Für die beiden anderen Bildungsmuster ist die Datenmenge zu gering für eine ausreichend genaue Bewertung. Die Autoren zeigen weiterhin, dass es auch unter den Nomen herausragend produktive Stämme gibt (z.B. *X-geschicht-lich*), was eine weitere Klassifizierung der Nomen nahe legt. Wie man an diesem Beispiel sieht, kann die qualitative Analyse von der quantitativen Analyse profitieren. Letztere fungiert sozusagen als Lackmustest für die Güte einer qualitativ begründeten Klassifizierung.

Anke Lüdeling, Stefan Evert und Ulrich Heid[24] zeigen aber auch, dass der automatischen Analyse von Korpora im Hinblick auf Anzahl und Häufigkeit von Wortbildungsmustern Grenzen gesetzt sind. Dies hängt mit der Fehleranfälligkeit der Analysemöglichkeiten zusammen, die eine manuelle Durchsicht der Daten beim heutigen Stand der Technik erforderlich machen. Probleme bereiten:

- Tippfehler in den Texten;
- Wörter, die zufällig mit der gleichen Zeichenkette wie das Suffix enden (z.B. *Balsam*, *Sesam*);

---

[22] Eine formale Beschreibung dieses als *Vocabulary Growth Curve* bezeichneten Phänomens gibt Baayen (2001). Siehe hierzu auch Baayen (2008).

[23] Vgl. Lüdeling und Evert (2003).

[24] Vgl. Lüdeling et al. (2000) und Evert und Lüdeling (2001).

- Wörter, die scheinbar eine Derivation sind, im Grunde aber eine Komposition mit einem früher derivierten Wort (z.B. *Kadavergehorsam* → *Kadaver+Gehorsam*, nicht jedoch → *Kadavergehor-sam*). Beide Fälle sind mit den heutigen Mitteln morphologischer Analyse nicht zu unterscheiden. So wurde z.B. *unversichtbar* gebildet durch Präfigierung von *versichtbar*; *befahrbar* wurde gebildet durch Suffigierung von *befahren*. Nur das letzte Wort ist relevant für die Wortbildung mit *-bar*[25].

Lüdeling und Evert zeigen das Potenzial, aber auch die Grenzen einer korpusgestützten Produktivitätsanalyse beim heutigen Stand der Technik[26]. Die Relevanz solcher Untersuchungen liegt in den folgenden Anwendungsgebieten:

- In der Lexikographie kann man sich bei unproduktiven Wortbildungselementen auf die Auflistung der wichtigsten lexikalischen Einheiten beschränken. Für produktive Wortbildungselemente ist der Ansatz eines eigenen Artikels zu erwägen, in dem die Verwendungsregularitäten erklärt werden sollten;
- im Fremdsprachunterricht spielt die Vermittlung der morphologischen und semantischen Regularitäten produktiver Wortbildungselemente eine wichtige Rolle. Es ist wahrscheinlich, dass Lerner Wörtern dieses Bildungstyps begegnen werden, die nicht im Wörterbuch stehen[27].

8.4 Syntax

In der Syntaxforschung kommen Korpora in verschiedener Hinsicht zum Einsatz. Sie werden als Quelle für authentische Beispiel herangezogen, die oftmals als Gegenbeispiele für einen in der Literatur vertretenen Standpunkt dienen sollen. Korpora bilden die Datengrundlage für die Erhebung von Frequenzangaben. Hierbei handelt es sich oft um den Vergleich von alternativen syntaktischen Realisierungsmöglichkeiten und das Korpus dient dazu, Kontextfaktoren einschließlich textexterner Metadaten zu identifizieren, die einen Einfluss auf die Variantenwahl haben. Diese Herangehensweise geht teilweise damit einher, dass nicht das Korpus selbst ausgewertet wird, sondern dass das Korpus nur als Samplinggrundlage für die eigentliche Datenbasis der Untersuchung herangezogen wird, die dann aus dem Korpus extrahiert und bei Bedarf weiter aufbereitet wird.

In welcher Form ein Korpus für syntaktische Forschungsfragen genutzt werden kann, hängt stark davon ab, welche Arten von Annotationen und Abfragewerkzeuge zur

---

25 Die Beispiele entstammen Evert und Lüdeling (2001).

26 Die Notwendigkeit manueller Intervention ist einer der Gründe, warum die Autoren für ihre *-lich*-Studie ein relativ kleines Korpus gewählt haben.

27 Korpusbasierte morphologische Analysen spielen auch in der Computerlinguistik und hier besonders in der Computerlexikographie eine Rolle. Korpusanalysen dienen hier dazu, das Regelhafte und das Idiosynkratische zu trennen: Alles, was nicht in Regeln gefasst werden kann, muss in Lexika beschrieben werden. Eine wichtige Rolle spielen hier die Arbeiten im Umfeld des morphologischen Lexikons *IMSLex*, vgl. Fitschen (2004). Wir können auf diesen Aspekt an dieser Stelle nicht näher eingehen und verweisen auf die computerlinguistische Fachliteratur.

Verfügung stehen. Wir haben bereits in Kapitel 4 gezeigt, dass es einen Unterschied ausmacht, ob man nur über Wortformen filtern kann oder ob auch Wortartenannotationen oder sogar weiterführende syntaktische Annotationen herangezogen werden können. In jedem Fall es sehr hilfreich, wenn man ein Suchwerkzeug mit regulären Ausdrücken verwenden kann[28]. Detmar Meurers und Stefan Müller[29] diskutieren eine Reihe von Fallbeispielen, in denen sie Korpusanfragen zu syntaktischen Phänomenen durchspielen. Sie erläutern anschaulich, wie man die linguistische Fragestellung in Konzepte der Korpusannotation übersetzen kann. Siehe hierzu auch Abschnitt 5.1.2 im vierten Kapitel und Abschnitt 6.2.1 in Kapitel 5.

Mangels verfügbarer Ressourcen haben Syntaktiker bisher oftmals nur mit wortbasierter Suche recherchiert z.B. Pittner (1999) oder Ehrich (2001) auf den IDS-Korpora. Das wird sich in Zukunft wahrscheinlich ändern, nachdem inzwischen die IDS-Korpora in weiten Teilen mit morpho-syntaktischen Annotationen angereichert sind, ebenso die DWDS-Korpora. Außerdem stehen inzwischen ja auch vollständig syntaktisch annotierte Baumbanken wie *TüBa-D/Z* und *TIGER* zur allgemeinen Verfügung.

Im Folgenden stellen wir zu den verschiedenen Nutzungsformen von Korpora in der syntaktischen Forschung stellvertretend ein paar Arbeiten vor. Ein Beispiel für die Suche nach (Gegen-)Beispielen, sind die Arbeiten von Stefan Müller zur mehrfachen Vorfeldbesetzung. Er verwendet für seine Recherchen die IDS-Korpora über die Online-Anfrage COSMAS, das Material, das auf den *DigiBib*-CDs[30] zur Verfügung steht, und die Tageszeitung *taz* (persönliche Auskunft). Das Ergebnis seiner Recherche sind Beispiele wie[31]:

(1) Öl ins Feuer goß gestern das Rote-Khmer-Radio: ...

Hier stehen zwei unabhängige Konstituenten im Vorfeld vor dem finiten Verb: *Öl* und *ins Feuer* vor *goß*. Die Belegsammlung dokumentiert die Natürlichkeit des Phänomens. Müller argumentiert, die Häufigkeit des Auftretens zeige, dass man die Daten, deren Existenz in der theoretischen Literatur wegen der Ungrammatikalität von Beispielen wie (2) teilweise bestritten wurde, nicht einfach ignorieren kann.

(2) * Maria Max gab ein Buch.

Müller selbst schlägt eine Analyse im Rahmen der *Head-Driven Phrase Structure Grammar* (HPSG) vor[32]. Die Datenbasis macht die Vielfalt des Phänomens deutlich und erlaubt es, Muster in den Daten festzustellen[33]. Die empirischen Daten helfen, Kontexteigenschaften zu identifizieren, die eine weitere Analyse unterstützen können.

In einer methodisch ähnlichen Arbeit untersucht Gabriele Knittka die Syntax und Pragmatik von NP-Aufsplittung im Deutschen (im Rahmen der sogenannten DP-Hypo-

---

[28] Siehe den Exkurs zu den *Regulären Ausdrücken* in Kapitel 4 auf S. 92.

[29] Vgl. Meurers (2005), Meurers und Müller (2008).

[30] Vgl. DigiBib: `https://www.hbz-nrw.de/angebote/digilink/`.

[31] Quelle: taz, 18.06.1997 – in alter Rechtschreibung.

[32] Konkret nimmt er an, dass die Konstituenten im Vorfeld durch ein abstraktes Verb lizenziert sind, vgl. Müller (2005).

[33] Siehe Müller (2003) und seine Belegsammlung auf `http://hpsg.fu-berlin.de/Software/TS/`.

these der Generativen Grammatik)[34]. Die Belege geschriebener Sprache stammen bei ihr aus verschiedenen Druckerzeugnissen, zusätzlich wertet sie aber auch ein kleines Korpus der gesprochenen Sprache aus.

Angelika Storrer[35] untersucht die Distribution von Nominalverbgefügen (NVG) wie *Unterricht erteilen*. Ein relativ allgemeines Verb (*erteilen*) tritt zusammen mit einer Nominalisierung als Objekt (*Unterricht*) in fester Wendung auf[36]. Storrer vergleicht die Verteilung der NVGs mit denen des jeweiligen Basisverbs (hier *unterrichten*). Motivation für diese Arbeit ist die immer wieder zu lesende Behauptung, dass die NVG nur eine phrasale Umschreibung des Basisverbes sei – und zudem ein schlechter Sprachstil. Anders als die bisher genannten Arbeiten wertet Storrer ein spezifisches Korpus aus, das DWDS-Kernkorpus. Sie analysiert die Belege zunächst qualitativ und untersucht dabei vergleichend das semantische und kombinatorische Potenzial von NVG und Basisverb, z.B. mögliche Selektionsrestriktionen oder Modifikationsmöglichkeiten am Basisverb und an der Nominalisierung. Letztere bietet eine Reihe von Optionen, die beim Basisverb nicht gegeben sind, wie die Modifikation durch bestimmte Adjektive, durch Relativsatz oder Spezifikator sowie bestimmte Koordinationsmöglichkeiten. Belege wie (3) im Kontrast mit dem konstruierten (4) können als Gegenbeispiel zur „Umschreibungsthese" gewertet werden.

(3) ... dem Krieg eine Absage erteilen.

(4) ?... dem Krieg absagen.

Eine zusätzliche quantitative Auswertung zur wechselseitigen Paraphrasierbarkeit ergibt, dass die Basisverben mehrdeutig (*polysem*), die entsprechenden NVGs hingegen spezifischer sind und meist nur eine der Bedeutungen des Basisverbs tragen. Die NVG erlaubt es demnach Ambiguitäten zu vermeiden. Zum Beispiel ist *unterrichten* ambig zwischen den Lesarten *mitteilen* und *lehren*, während *Unterricht erteilen* nur die eine Bedeutung hat. Das Fazit der Studie ist, dass Nominalverbgefüge keine „semantischen Dubletten" des Basisverbs sind – die oben erwähnte Stilfrage stellt sich damit nicht. Storrers Arbeit leitet direkt zum zweiten Verwendungstyp über, dem der Frequenzanalyse.

Die im Folgenden dargestellten Arbeiten erheben Frequenzdaten auf einem syntaktisch annotierten Korpus. Sie sind beide an der Distribution von Relativsätzen interessiert und verbinden die Untersuchung der Korpusfrequenz mit psycholinguistischen Experimenten. In der ersten Arbeit untersuchen Uszkoreit et al.[37], welche Faktoren einen Einfluss darauf haben, ob ein Relativsatz adjazent, d.h. direkt benachbart, zu seinem Bezugsnomen steht (5) oder extraponiert im Nachfeld auftritt (6).

(5) Er hat [das Buch, [das er gestern erst gekauft hat],] heute gelesen.

(6) Er hat [das Buch] heute gelesen, [das er gestern erst gekauft hat].

---

[34] Vgl. Kniffka (1996).

[35] Vgl. Storrer (2006a).

[36] Die Klasse der Nominalverbgefüge ist in sich nicht homogen. Storrer (2006b) differenziert hier weiter und stellt einen korpusbasierten Vergleich von zwei Subklassen vor.

[37] Vgl. Uszkoreit et al. (1998).

Die Studie basiert auf einer Vorstufe des NEGRA-Korpus mit 12 000 vollständig syntaktisch annotierten Sätzen, welches sich aber als zu klein erwies, so dass die Autoren auf ein weiteres Korpus zurückgreifen. Die Untersuchung konnte so auf einer Textbasis von 1 Millionen Wörtern durchgeführt werden[38]. Das Ergebnis der quantitativen Studie legt eine performanzorientierte Erklärung der Distribution nahe. Bestimmend sind die Faktoren *Distanz* (zwischen Bezugsnomen und potenzieller extraponierter Position) und *Länge* (Gewicht des Relativsatzes in Wortanzahl). Eine ähnliche Auswertung, diesmal auf dem kompletten NEGRA-Korpus, wird von Schade et al.[39] durchgeführt. Sie suchen nach geschachtelten Relativsätzen in der geschriebenen Sprache und finden Beispiele wie (Klammerung wurde hinzugefügt):

(7) Er hat jene Heiterkeit, [die ein Tierlehrer, [der an sich auf Pferdedressuren geeicht ist], braucht], um auch ein so spaßiges Spektakel wie den „Schweizer Bergbauernhof" durchzustehen.

Um einen Eindruck von der spontanen Produktion zu bekommen, werten sie auch die Verbmobil-Baumbank zur gesprochenen Sprache aus[40]. Dort finden sie keine geschachtelten Relativsätze, sondern nebengeordnete Strukturen wie (Klammerung wiederum hinzugefügt):

(8) Ja, also erstmal zum Hotel: Da haben wir noch drei verschiedene Hotels, [die wir Ihnen anbieten können], [die noch Zimmer frei haben].

Die beiden Korpusstudien verwenden Schade et al. als Ausgangsbasis für ihre weiterführenden psycholinguistischen Experimente zur Relativsatzperzeption.

Ebenfalls eine Triangulation von Methoden setzt Amir Zeldes[41] in seiner Untersuchung des deadjektivischen, präpositional verwendeten *voller* in Konstruktionen wie *eine Wanne voller Wasser* ein. Bei diesem besonderen Wort handelt sich um einen syntaktischen Einzelgänger, d.h. ein Wort das sich nicht ohne weiteres in die klassischen Wortarten eingliedern lässt. An den Konstruktionen mit *voller* ist auffallend, dass es selbst für Muttersprachler sehr schwierig ist, den Kasus des artikellosen Objekts von *voller* eindeutig zu bestimmen. Wenn ein Adjektiv eingefügt wird wie in Beispiel (9) sind die Intuitionen besser, aber auch hier streiten sich die Geister[42].

(9) eine Wanne voller warmem$_{DAT}$ Wasser

Zeldes wertet das wortartengetaggte deWaC Web-Korpus[43] aus und extrahiert etwa 21 000 Kandidaten für *voller*-Konstruktionen. Die Kandidatenliste filtert er weiter und

---

[38] Das zweite Korpus ist nur POS annotiert und erfordert, wie die Autoren bemerken, viel zeitaufwändige Handarbeit in der Auswertung.

[39] Vgl. Schade et al. (2003).

[40] Die Verbmobil-Baumbank ist 2005 als *TüBa-D/S* veröffentlicht worden.

[41] Vgl. Zeldes (erscheint).

[42] Wenn Sie die Konstruktion googlen, finden Sie auch Belege für *voller warmen Wasser* oder *voller warmen Wassers*. Es sei jedem Leser selbst überlassen, eigene Akzeptabilitätsurteile zu fällen.

[43] deWaC umfasst 1,63 Milliarden Token, vgl. Baroni et al. (2009).

taggt sie mit dem RFTagger[44], der anders als der im deWaC eingesetzte TreeTagger zusätzlich zu den Wortarten auch morphologische Informationen wie Kasus, Numerus und Genus annotiert. Der quantitativen Studie liegt schlussendlich ein Datenset von etwa 20 500 morphologisch getaggten *voller*-Belegen zugrunde, deren Objektphrase in mehr als 87% der Fälle aus einem nicht weiter modifizierten Nomen besteht.

Zusätzlich zu den Webdaten untersucht Zeldes ein kleines Korpus mit geschriebenen Spracherwerbsdaten von Schulkindern[45]. Diese Ergebnisse sind wegen der kleinen Datengrundlage nur mehr anekdotisch, zeigen aber, dass die Kasuszuweisung in der *voller*-Konstruktion für Schulkinder in der vierten Klasse ein Problem darstellt.

Auf die introspektiven Daten, die Zeldes im Zusammenhang mit den quantitativen Ergebnissen diskutiert, wollen wir hier nicht weiter eingehen, sondern seine Interpretation der Daten zusammenfassen: *Voller* ist eine Art Präposition, weist aber zwei Besonderheiten in Bezug auf die Nominalphrase, die es regiert, auf: Die Nominalphrase darf keinen Artikel enthalten und man beobachtet eine Art differenzielle Objektmarkierung[46], die sonst im Deutschen nicht attestiert ist: *voller* weist unterschiedlichen Kasus mit unterschiedlichen Frequenzen in Abhängigkeit von Eigenschaften der Objektphrase selbst zu. Eine große Rolle spielt dabei der Numerus (Dativ Singular vs. Genitiv Plural), teilweise ist auch die morphologische Klasse relevant und der Umstand, ob das Kopfnomen von einem Adjektiv modifiziert wird oder nicht. Abschließend formalisiert Zeldes seine Analyse im Rahmen der Sign-Based Construction Grammar[47].

Eine semantisch motivierte Konstruktion steht im Mittelpunkt von einer Untersuchung von Timm Lichte, die wir Ihnen hier ebenfalls vorstellen wollen[48]. Lichte arbeitet mit einem rekursiv gechunkten Korpus, der TüPP-D/Z. Er verwendet 2,7 Millionen Sätze des Gesamtkorpus, um automatisch sogenannte Negative Polaritätselemente (NPI)[49] zu identifizieren. NPIs sind Ausdrücke, die nur im Umfeld von bestimmten negativen Ausdrücken und Fragekontexten lizenziert sind wie *ganz geheuer* in dem Satz *Das ist mir nicht ganz geheuer*. Lichte legt die Annahme zu Grunde, dass sich NPIs und ihre Lizenzierer wie Kollokationen verhalten. Außer der Menge der Lizenzierer gelten alle anderen Lemmata des Korpus als potentielle NPIs[50]. Sein System erstellt eine Rangliste der Lemmata, die manuell überprüft werden muss. Unter den obersten 20 Kandidaten findet man schöne Beispiele wie *verdenken*, *unversucht*, *umhin* oder *lumpen*. Lichte zeigt auch auf, wie seine Methode auf Mehrwort-NPIs erweitert werden kann. In einem Experiment dazu erhält er Kandidaten wie *unversucht lassen*, *ganz geheuer*, *umhin zu kommen* oder *lumpen lassen*.

---

[44] RFTagger: vgl. Schmid und Laws (2008).

[45] Das KESS-Korpus, Kompetenzen und Einstellungen von Schülerinnen und Schülern, wurde vom Landesinstitut für Lehrerbildung und Schulentwicklung in Hamburg erhoben.

[46] Vgl. Bossong, Georg (1985).

[47] Vgl. Boas und Sag (2012).

[48] Vgl. Lichte (2005).

[49] Auf Englisch ‚Negative Polarity Item', daher die Abkürzung NPI.

[50] Lichte beschränkt die Untersuchung auf Lemmata, die häufiger als 40 mal im Korpus vorkommen. Er erhält damit eine Ausgangsmenge von fast 35 000 Lemmata.

8.5 Lexikologie und Lexikographie

Der Nutzen von Korpora für die Lexikographie ist vielfältig, was an anderer Stelle ausführlich beschrieben wird[51]. Wir wollen uns hier auf eine Zusammenfassung aus der Sicht des lexikographischen Prozesses und auf einige Felder beschränken, die auch für das Deutsche gut bearbeitet wurden.

Aus der Sicht des lexikographischen Prozesses[52] werden Korpora in den folgenden Phasen konsultiert:

- Bei der Wörterbuchplanung, besonders bei der Finanzplanung, spielen die Existenz und die Verfügbarkeit von Korpora für den durch das Wörterbuch zu beschreibenden Gegenstand eine Rolle. Wichtig sind auch die Werkzeuge, die die für die Lexikographen relevanten Informationen aus den Korpora extrahieren und präsentieren. Hier ist möglicherweise Entwicklungs- und Anpassungsarbeit notwendig.
- Korpora können wichtige Hinweise für die Lemmaauswahl geben. So kann die Häufigkeit, mit der eine lexikalische Einheit in einem Korpus vorkommt, darüber entscheiden, ob sie in die Lemmaliste eines Wörterbuchs aufgenommen wird oder nicht[53].
- Den Hauptteil lexikographischer Arbeit bildet das Erstellen der Wörterbuchartikel zu den Lemmata. Bei einem allgemeinsprachlichen Standardwörterbuch müssen die lexikalischen Zeichen auf allen linguistischen Ebenen beschrieben werden. Hierfür bilden Korpora eine Informationsquelle[54].
  Betrachten wir ein Beispiel. Es muss beschrieben werden, ob bestimmte Verben, die mentale Zustände ausdrücken – *wissen, glauben, meinen* etc.
  - mit *dass*-Sätzen und *ob*-Sätzen als Ergänzung verwendet werden können; wenn dies der Fall ist
  - welches, wenn beide Ergänzungen möglich sind, die häufigere Variante ist oder ob eine der beiden Varianten sehr selten ist, und weiter
  - ob die Verwendung der Ergänzungen auf bestimmte Kontexte beschränkt ist, z.B. negative Kontexte oder bestimmte Zeitformen des Verbs:

  (10) *Ich weiß, ob das geht.

  (11) Ich weiß *nicht*, ob das geht.

  (12) *Er wusste, ob das geht.

  (13) Er *wird* schon wissen, ob das geht[55].

---

51 Vgl. Engelberg und Lemnitzer (2009), Wiegand (1998) und die dort erwähnte Literatur sowie, für das Englische, Ooi (1998).

52 Vgl. hierzu vor allem Kapitel 6 in Engelberg und Lemnitzer (2009).

53 Ausführlich hierzu Geyken und Lemnitzer (2012).

54 Lemnitzer und Geyken (Lemnitzer und Geyken (2014) zeigen die Möglichkeiten aber auch Grenzen der Extraktion von lexikographischen Angaben aller Art aus Textkorpora.

55 Wir empfehlen Ihnen, in einem Wörterbuch ihrer Wahl nachzuschlagen und zu prüfen, ob Sie auf die Fragen, die wir hier gestellt haben, eine Antwort finden. Wenn Sie Muttersprachler sind, versetzen Sie sich in die Situation eines Nichtmuttersprachlers, der diese Verben korrekt verwenden möchte. Oder machen Sie den Test mit einem Wörterbuch einer anderen Sprache.

Diese subtilen Unterscheidungen können am besten durch die gründliche Analyse eines Textkorpus ermittelt werden.

- Korpora stellen eine wichtige Quelle von Verwendungsbeispielen dar. Lexikographen können auf Grund ihrer Sprachkompetenz zwar Beispiele erfinden, es hat sich aber erwiesen, dass diese bei weitem nicht an die Qualität von Korpusbelegen heranreichen[56].
- Die Häufigkeit ihrer Verwendung kann ein wichtiges Kriterium für die Anordnung von Lesarten in einem Artikel für ein sprachliches Zeichen sein. Vor allem in Lernerwörterbüchern sollte das Häufige vor dem Seltenen erscheinen oder das Seltene sogar unerwähnt bleiben, je nach Umfang des Wörterbuchs.
- Ein wichtiger Aspekt der Verwendung lexikalischer Zeichen ist ihre Verwendung in typischen Kotexten. Manche lexikalischen Zeichen tauchen in nur einem oder sehr wenigen Kotexten auf (z.B. *Hehl, fackeln*), viele lexikalische Zeichen treten typischerweise mit einer kleinen Anzahl anderer lexikalischer Zeichen auf und bilden mit diesen Kollokationen oder idiomatische Wendungen (typische Begleiter von *hart* sind z.B.: *Bandagen, Droge, Leben, Währung*). Statistische Verfahren, auf großen Korpora angewendet, geben Auskunft über diese typischen Paarungen. Auch hier sind Korpora der sprachlichen Intuition – selbst der von den erfahrensten Lexikographen – überlegen.
- In den Produktionsphasen nach der Erstellung der Wörterbuchartikel – Korrektur und Drucklegung – spielen Korpora naturgemäß eine geringe Rolle. Einzelne Entscheidungen in der Korrekturphase können bei Bedarf an Korpora überprüft werden. In der Phase der Materialsammlung zwischen zwei Auflagen eines Wörterbuchs kommt Texten, die nach der Drucklegung der letzten Auflage erschienen sind, wieder eine größere Bedeutung zu.

Die Werkzeuge, die Lexikographen typischerweise für diese Arbeit verwenden, sind Programme für die quantitative Analyse von Korpora, um z.B. die Verwendungshäufigkeit bestimmter lexikalischer Zeichen – insgesamt oder in bestimmten Lesarten – oder typische Kombinationen sprachlicher Zeichen zu ermitteln. Des Weiteren werden Programme verwendet, die für ein bestimmtes lexikalisches Zeichen alle Vorkommenskontexte in einer vom Lexikographen festlegbaren Anordnung präsentieren[57]. Die Kombination dieser Werkzeuge hilft, aus dem Meer der Texte durch Auswahl und Filterung der Daten den Lexikographen die Informationen zu liefern, die sie für ihr Handwerk der lexikalischen Beschreibung benötigen[58].

Wir werden uns im Folgenden auf drei Felder konzentrieren, auf denen die germanistische Korpuslinguistik bereits einige Erfolge erzielen, d.h. interessante und relevante Ergebnisse zu Tage fördern konnte. Dies sind die Lexikonbereiche der Neologismen und Anglizismen sowie die Kombination einzelner lexikalischer Zeichen in Kollokationen

---

[56] Luise Pusch hat eine lesenswerte Satire geschrieben, für die sie reichlich Beispiele der von den Duden-Redakteuren produzierten Belegprosa verwendet, vgl. Pusch (1984).

[57] Diese Werkzeuge präsentieren ‚Keywords in Context', und werden deshalb KWIC-Tools genannt, die Daten, die sie erzeugen, *Konkordanzen*.

[58] Ein Desiderat sind allerdings immer noch Werkzeuge, die automatisch die Belege auswählen, in denen ein Schlüsselwort in einer bestimmten Lesart verwendet wird. Dies ist ein Forschungsgegenstand der Computerlinguistik.

und festen Wendungen. Als Spezialfall von Kollokationen werden wir im Anschluss auf Kombinationen von Modalpartikeln eingehen.

8.5.1 Neologismen

Im weitesten Sinne sind Neologismen sprachliche Zeichen, also Wörter, Bedeutungen und Wendungen, die zu einem bestimmten Zeitpunkt von den Sprechern, die sie verwenden, als neu empfunden werden.

Neologismen können von ihrer Form her unterteilt werden in Neulexeme und Neubedeutungen. Das Wort *Podcast* ist vor nicht allzu langer Zeit als ein Neulexem in den deutschen Sprachgebrauch aufgenommen worden, da es diese Wortform im Deutschen Lexikon bisher nicht gab[59]. Das Wort *Maus* hingegen erhielt in den frühen siebziger Jahren eine Neubedeutung, es bezeichnet seitdem ein Steuergerät am Computer.

Neologismen können weiterhin an Hand des Grades ihrer Lexikalisierung und ihrer Integration in den deutschen Sprachgebrauch unterschieden werden. Danach bezeichnen Neologismen im engeren Sinn Wörter, die weitgehend lexikalisiert sind. Sie werden relativ häufig und bereits über einen längeren Zeitraum verwendet und in die Neuauflagen allgemeinsprachlicher Wörterbücher aufgenommen. Hierzu gehört sicher das Verb *simsen* (=eine SMS verschicken). Daneben gibt es die Gelegenheitsbildungen, die nur ein oder wenige Male verwendet werden, danach wieder in Vergessenheit geraten und auch nicht in Wörterbücher aufgenommen werden. Ein Beispiel hierfür ist das Wort *semimerkelig* (womit eine Frisur im Stil von Angela Merkel bezeichnet wurde). Diese sogenannten *Okkasionalismen* sind von der Lexikographie und Lexikologie lange Zeit als uninteressant abgetan worden. Sie bieten aber für die Wortbildungsforschung und für die Lexikographie interessantes Material[60]. Entlang dieser letzten Unterscheidung haben sich zwei Formen der Neologismenlexikographie herausgebildet:

- Die *aktuelle Neologismenlexikographie* sammelt und archiviert Wörter vom ersten Augenblick ihres Erscheinens an. Diese Sammlungen enthalten zwangsläufig viele Okkasionalismen, da zum Zeitpunkt des ersten Erscheinens eines Wortes nicht vorhergesagt werden kann, ob dieses Wort sich im Gebrauch etablieren wird. Erfahrene Lexikographen können lediglich gute Voraussagen über die Entwicklung eines Wortes treffen. Ein Beispiel für die aktuelle Neologismenlexikographie ist die *Wortwarte*[61].
- Die *retrospektive Neologismenlexikographie* sammelt und beschreibt in Spezialwörterbüchern dieses Lemmatyps die Wörter, die im Beschreibungszeitraum aufgekommen sind und sich bereits etabliert haben. Ein Beispiel hierfür sind die am Institut für Deutsche Sprache erschienen Wörterbücher zum neuen Wortschatz[62]. Dementsprechend wird hier der Begriff *Neologismus* im engeren Sinn verwendet.

Korpusdaten haben in der Neologismenforschung und -lexikographie die folgenden Funktionen:

---

[59] *Podcast* bezeichnet die meist private Distribution von Hörbeiträgen, im Stile eines Radiosenders, über das World Wide Web.

[60] Vgl. hierzu Peschel (2002), Tomášiková (2008) und Lemnitzer (2013).

[61] Im WWW unter der Adresse `www.wortwarte.de` erreichbar.

[62] Vgl. Herberg et al. (2004) und Steffens und al Wadi (2013).

- Bei regelmäßiger Beobachtung zum Beispiel der Tagespresse lässt sich mit einiger Sicherheit feststellen, wann ein Wort (in einer bestimmten Bedeutung) zum ersten Mal verwendet wurde (Erstbeleg).
- Die quantitative Auswertung eines größeren Korpus, das den Sprachgebrauch eines bestimmten Zeitraums repräsentiert, ergibt, welche Wörter ausreichend oft belegt sind, so dass man von einem etablierten Wort, also einem Neologismus im engeren Sinn sprechen kann. Es lassen sich auf diese Weise auch Profile der Gebrauchshäufigkeit von Neologismen, die schon länger im Gebrauch sind, ermitteln und auch feststellen, welche Neologismen nach einer gewissen Zeit wieder außer Gebrauch kommen[63].
- Anhand eines zeitlich gegliederten Korpus lässt sich auch ermitteln, welche Wortbildungselemente eine wachsende Rolle bei der Bildung neuer Wörter spielen. So ist z.B. das Präfix *Cyber-* erst seit Ende des letzten Jahrzehnts in Verwendung und gehört seitdem zu den produktiven Wortbildungselementen.
- In Korpora belegte Verwendungsgewohnheiten geben Auskunft über sich verfestigende Eigenschaften des Gebrauchs, z.B. die Zuordnung eines Genus zu einem aus dem Englischen entlehnten Wort.
- Schließlich liefern Korpora Belege, die als Vorlagen für den Erwerb des normgerechten Gebrauchs eines neuen Wortes wichtig sind.

Linguistische und lexikographische Neologismus-Forschung ist also ohne die Analyse authentischer Sprachdaten unmöglich. Für lange Zeit war die manuelle Analyse und Auswertung von Printwerken die einzig machbare Arbeitsmethode, und vor allem in der Wörterbucherstellung werden neue Wörter noch heute überwiegend auf diese Art gesammelt. Es gibt aber Projekte, in denen digitalisierte Korpora für diese Zwecke genutzt werden.

Ein Beispiel hierfür ist die *Wortwarte*. Seit Ende 2000 werden täglich die Online-Ausgaben mehrerer Tages- und Wochenzeitungen ausgewertet. Die Wörter dieser Texte werden mit der Wortliste eines Referenzkorpus abgeglichen. Die nach diesem Abgleich übrig gebliebenen Wörter werden täglich durchgesehen und im Durchschnitt 15 neue Wörter ausgewählt, beschrieben und mit einem Beleg aus der Fundstelle versehen. Neben dem online zugänglichen Wörterbuch mit mittlerweile über 60 000 Einträgen stehen alle Wortlisten zur Verfügung. Mit diesen Daten lassen sich z.B. Aussagen über Tendenzen der Wortbildung treffen[64]. Auch in diesem Projekt wird mit einem weiten Begriff von *Neologismus* gearbeitet, der auch Gelegenheitsbildungen umfasst. Zweitens wird in diesem Projekt, und dies ist ein neuer Ansatz, versucht, das Web, genauer: einen kleinen Ausschnitt daraus, als kontinuierliche Quelle aktueller Sprachdaten zu nutzen.

Ein größeres Spezialwörterbuch des Lemmatyps Neologismen, das der retrospektiven Neologismenlexikographie verpflichtet ist, bildet die vom Institut für deutsche Sprache herausgegebene Sammlung *Neuer Wortschatz. Neologismen der 90er Jahre* von Die-

---

[63] Steffens und al-Wadi nennen dies Zeitverlaufgrafik, vgl. Steffens und al Wadi (2013), S. XXIVf.

[64] Die Einträge sind auf der Website der Wortwarte, `www.wortwarte.de`, veröffentlicht, welche täglich aktualisiert wird. Auf der Website befinden sich auch weitere Informationen zum Projekt. Die Wortlisten können beim Autor angefordert werden.

ter Herberg, Michael Kinne und Doris Steffens[65]. Bei der Erstellung dieses Wörterbuchs wurde mit einem engeren Neologismusbegriff gearbeitet. Gegenstand des Wörterbuchs sind die Neuwörter und Neubedeutungen, die

> in den 90er Jahren des 20. Jahrhunderts in der deutschen Allgemeinsprache aufgekommen sind, sich darin ausgebreitet haben, als sprachliche Norm allgemein akzeptiert und in diesem Jahrzehnt von der Mehrheit der deutschen Sprachbenutzer über eine gewisse Zeit hin als neu empfunden wurden. (Herberg et al. 2004, S. XXIII)[66]

Das Erscheinungsjahr des Wörterbuchs, 2004, und das der Fortsetzung, 2013, zeugen, dass die Autoren zwar zeitlich relativ nah an ihrem Beschreibungsgegenstand sind, aber doch weit genug entfernt, um den Prozess der Lexikalisierung aus der Rückschau beobachten zu können. Als Primärquelle des Werks diente ein Teil der IDS-Korpora, das Texte des untersuchten Zeitraums umfasst. Dazu kam eine Wortkartei mit ca. 10 000 Einträgen (S. XVI f.). Inwiefern sich ein solches lemmabezogenes Spezialwörterbuch neben aktuellen allgemeinsprachlichen Wörterbüchern, vor allem dem Rechtschreibduden, etablieren wird, bleibt abzuwarten. Die beiden Bücher und die begleitende Online-Version[67] sind jedenfalls eine interessante Quelle für die Fremdvermittlung bei fortgeschrittenen Lernern. Vielleicht ergeben sich aus dieser konsequent korpusbezogenen Arbeit auch Impulse für die traditionelle Lexikographie des Deutschen und deren Produkte.

Schließlich sollen noch einige Spezialarbeiten zu Neologismen aus linguistischer Sicht, und hier vor allem die Beiträge von Hilke Elsen zu Neologismen in einigen Varitäten des Deutschen, erwähnt werden[68].

Mit den beschriebenen Projekten hat sich eine linguistische und lexikographische Praxis der Analyse von Neologismen auch des Deutschen etabliert. Neu sind vor allem die Nutzung des World Wide Web als Datenquelle und die stärkere Berücksichtigung von Okkasionalismen.

8.5.2 Anglizismen

Anglizismen sind ein weiterer markierter Bereich des deutschen Wortschatzes. Unter dem Begriff Anglizismus versteht man alle aus dem Sprachkontakt einer Sprache mit dem Englischen resultierenden Phänomene der Entlehnung und der Beeinflussung des Sprachsystems der (in unserem Fall deutschen) Zielsprache[69]. Aus vielerlei Gründen ist das Englische nach 1945 zur stärksten Gebersprache im linguistischen Kontakt geworden. Aus dem britischen und vor allem dem amerikanischen Englisch entlehnte lexikalische Einheiten bilden einen nicht zu vernachlässigenden Teil des Vokabulars der deutschen Sprache. Die Integration dieser Wörter ist dabei in das System der deutschen

---

[65] Herberg et al. (2004). Mit Steffens und al Wadi (2013) ist mittlerweile eine Fortsetzung erschienen, deren Darstellungszeitraum die Jahre 2001–2010 ist. Ein weiteres solches Wörterbuch, auf das hier nur kurz hingewiesen werden kann, ist Quasthoff (2007).

[66] Besonders das letzte Kriterium steht auf empirisch schwachen Füßen. Es ist zu vermuten, dass das Sprachgefühl der Autoren hier repräsentativ für das Sprachgefühl aller Sprachbenutzer gesetzt wird.

[67] Unter `www.owid.de` am Institut für Deutsche Sprache verfügbar.

[68] Vgl. Elsen (2002), Elsen (2004) und Elsen und Dzikowicz (2005).

[69] Vgl. Bartsch (2002), S. 312.

Sprache ist dabei mehr oder weniger fortgeschritten. Anglizismen stellen das System und vor allem den Gebrauch der deutschen Sprache vor besondere Schwierigkeiten.

- Orthographisch weicht die Norm der Getrennt- und Zusammenschreibung sowie der Bindestrichschreibung von der englischen Norm und orthographischen Praxis ab[70].
- Die Aussprache kann sich eher am englischen Original orientieren (z.B. *Banker* [bæŋkə] anstatt [baŋkə] oder *kiten* [kaitŋ] anstatt [kiːtn̩]) oder am phonologischen System des Deutschen (z.B. bei *Download* wird die zweite Silbe eher als [loːt] gesprochen mit deutscher Auslautverhärtung anstatt des ursprünglichen [loːd]).
- Morphologisch ergeben sich Probleme bei der Genitiv- und der Pluralbildung (*Flyer* → *?Flyers* oder *?Flyer*) und der Konjugation (*?geuploaded, ?upgeloaded*)[71].
- Die größten Probleme entstehen beim Genus, das im Englischen nicht festgelegt ist (der / die / das *Engine, Toolbar, Airbag*?). In einer Untersuchung zu diesem Thema kommt Rudolf-Josef Fischer, der u.a. auch Sprecherurteile einbezieht, zu dem Ergebnis, dass keine Kombination der in der Literatur zur Genuszuweisung bei (neuen) Substantiven diskutierten Prinzipien dazu in der Lage ist, diesen Prozess vollständig zu erklären[72].
- Grammatisch ergeben sich die geringsten Probleme, da die Systeme sich hier sehr ähneln (heißt es *Aktien traden* oder *mit Aktien traden*, letzteres in Analogie zu *handeln*?).
- Weiterhin bringen Anglizismen Unsicherheiten in der Verwendung mit sich – *Searchengine* wird man wahrscheinlich nicht im Gespräch mit der Großmutter verwenden und *abchillen* nicht im Gespräch mit dem Chef.

Wie man sieht, müssen die Verwendungsbedingungen von entlehnten Wörtern erst im Prozess der Entlehnung ausgehandelt werden, besonders dort, wo sie in der Gebersprache nicht ausreichend spezifiziert sind[73]. Die Integration in das sprachliche System des Deutschen kann unterschiedlich weit fortschreiten (vgl. *Majonäse* oder *Kode*, im Gegensatz dazu ist der Ausdruck *Computer* kaum integriert). Sie wird von Normen wie etwa der zur Rechtschreibung gesteuert, und die Aufnahme eines Anglizismus in die Wörterbücher des Deutschen geht mit Festlegungen der Verwendungsnorm auf den verschiedenen linguistischen Ebenen einher.

Anglizismen werden bevorzugt in drei Wörterbuchtypen aufgenommen:

- Spezialwörterbücher des Lemmatyps Anglizismus. Hier ist vor allem das sprachdokumentarische *Wörterbuch der Anglizismen* von Carstensen und Busse zu nennen[74].

---

70 Vgl. hierzu, aus dem Blickwinkel der alten Rechtschreibnorm, Augst (1992).

71 Der Rechtschreibduden schlägt als Norm für das Perfektpartizip des Lexems *e-mailen* die Form *ge-mailt* vor. Ähnlich ungewöhnlich nimmt sich die immerhin im DWDS-Korpus mehrfach belegte Form *ge-e-mailt* aus. Auch dies ist ein Beispiel für die Schwierigkeiten bei der (orthographischen) Integration englischer Lehnwörter.

72 Vgl. Fischer (2005).

73 Die nicht vorhandene Genusmarkierung bei englischen Nomen ist hierfür ein Beispiel.

74 Vgl. Carstensen und Busse (1993). Die lexikographische Arbeit stützt sich auf das Paderborner Korpus, im Wesentlichen eine Belegsammlung, sowie die Korpora, die Mitte der achtziger Jahre am Institut für deutsche Sprache zur Verfügung standen, vgl. Carstensen und Busse (1993), S. 47–53.

Es gibt aber auch einige sprachpuristisch ausgerichtete Werke auf diesem Regalbrett, z.B. das *Wörterbuch überflüssiger Anglizismen* von Bartzsch[75].

- Fremdwörterbücher, in denen die aus anderen Sprachen entlehnten oder aus dem Griechischen und Lateinischen überkommenen lexikalischen Einheiten versammelt sind, deren Gebrauch in der Alltagssprache weniger üblich ist (z.B. *Parallaxe, Chintz*).
- Allgemeinsprachliche Standardwörterbücher wie das Duden Universalwörterbuch oder Spezialwörterbücher z.B. zur Rechtschreibung.

Normunsicherheit besteht vor allem bei Wörtern, die noch nicht in Wörterbüchern registriert sind. Im Prinzip sollten hier die generellen orthographischen und grammatischen Normen des Deutschen hinreichend präzise Richtlinien für den Gebrauch geben. Augst zeigt jedoch, dass zumindest die Regeln der (alten) Rechtschreibung nicht ausreichen und selbst in den Wörterbüchern bei einzelnen lexikalischen Einheiten inkonsequent angewendet wurden[76]. Auch die Regeln der reformierten Rechtschreibung erleichtern es nicht, die korrekte Schreibung eines Anglizismus zu erschließen, wie Jürgen Dittmann und Christian Zitzke zeigen[77]. Die Autoren zeigen weiterhin in einer korpusbasierten Studie, dass in einigen Bereichen der Sprachgebrauch deutlich von den Normen, der offiziellen wie auch der der Nachrichtenagenturen, abweicht[78]:

- Bei rein englischen Komposita dominiert die Getrenntschreibung, eine deutliche Abweichung von beiden Normen (z.B. *Key Accounter, Call Center*);
- bei den Mischkomposita mit englischen und deutschen Bestandteilen dominiert die normgerechte Zusammenschreibung, gefolgt von der Bindestrichschreibung, die von der Norm zumindest toleriert wird (z.B. *Produktmanager, Softwareentwicklungsmethoden*); mehrgliedrige Komposita mit einem Funktionswort als Bestandteil (z.B. *Business-to-Business*) werden ebenfalls meist normkonform mit Bindestrich gebildet und durchgekoppelt, es bestehen hier aber große Unsicherheiten hinsichtlich der Klein-/Großschreibung der einzelnen Bestandteile – nominale Bestandteile müssen hier groß-, nicht-nominale Bestandteile kleingeschrieben werden.

Die Autoren beobachten, dass erstens die Anlehnung an den Gebrauch in der Quellsprache (bei den rein englischen Komposita), zweitens die Vertrautheit der einzelnen fremdsprachlichen Elemente und drittens die Länge des Gesamtkompositums eine Rolle bei der Wahl der Schreibweise (getrennt, mit Bindestrich oder zusammen) spielen. Eine Ausrichtung an der Norm dürfte eher zufällig sein, zumal, wie die Autoren im ersten Teil ihrer Arbeit zeigen, sich aus der Norm nur schwer Gebrauchs-Richtlinien ableiten

---

75 Vgl. Bartzsch (2004).

76 Vgl. Augst (1992), u.a. S. 58.

77 Vgl. Dittmann und Zitzke (2000), vor allem S. 70–76. Dittmann und Zitzke untersuchen in dieser Hinsicht sowohl die offiziellen Regeln als auch die Richtlinien der Nachrichtenagenturen.

78 Die Autoren verwenden als Datenbasis die Stellenanzeigen aus der Frankfurter Allgemeinen Zeitung, der Süddeutschen Zeitung und der Welt vom 24. April 1999 und der Neuen Zürcher Zeitung vom 5. Mai 1999. Ihre quantitative Auswertung stützen sie auf die 4225 Vorkommen von Anglizismen in den beiden erstgenannten Zeitungen, vgl. Dittmann und Zitzke (2000), S. 77.

lassen. Dittmann und Zitzke belegen all ihre Befunde mit exakten Zahlen, die sie durch Auszählung der Vorkommen in ihrem Korpus ermitteln.

Eine neuere und methodisch interessante, weil konsequent korpusgestützte Arbeit zu diesem Thema hat Peter Eisenberg vorgelegt[79]. Er bezieht sich in seinen Untersuchungen auf zwei ausgewählte Zeitscheiben aus dem Kernkorpus des DWDS, nämlich die Jahre 1905–1914 und 1995–2014 (62). Die Wortformen dieser Korpustexte mit einem Umfang von jeweils etwa 10 Millionen Token wurden lemmatisiert, was für jede Zeitscheibe eine Lemmaliste von knapp 400 000 Einträgen ergibt (63). Die Anglizismen in dieser Liste wurden nach einer Arbeitsdefinition des Konzepts *Anglizismus* (69ff.) manuell annotiert (64). Damit liegt eine quantitativ wie qualitativ auswertbare Datenbank für zwei jeweils zehnjährige und weit auseinander liegende Perioden des 20. und frühen 21. Jahrhunderts vor. In den folgenden Abschnitten der Arbeit werden die Daten quantitativ und qualitativ nach den üblichen Beschreibungsebenen wie Orthographie, Phonologie und Morphologie untersucht (75–114). Eisenberg kommt dabei zu den folgenden, hier nur kurz zusammengefassten Schlüssen: a) Anglizismen stehen im Deutschen unter starkem Anpassungsdruck der deutschen Kerngrammatik. Im Gegenzug haben sie bisher die deutsche Kerngrammatik kaum beeinflusst; b) im intensiven Kontakt des Deutschen mit dem (britischen und amerikanischen) Englisch werden vor allem dort Entlehnungen gemacht, wo ein Benennungsbedarf für neue Gegenstände und Sachverhalte besteht; c) Kritik ist am Gebrauch von Anglizismen, aber auch von Wörtern anderer Arten, gerechtfertigt, wo dieser eigentlich einen Missbrauch darstellt, weil er unakzeptablen Zwecken dient. Nur hier hat die Sprachgebrauchskritik, auch die, die sich auf Anglizismen bezieht, eine gewisse Berechtigung (115).

8.5.3 Kollokationen und Phraseme

Als Kollokation wird das gemeinsame Vorkommen zweier sprachlicher Zeichen miteinander bezeichnet. Ein Element einer Kollokation tritt im Umfeld des anderen Teils auf. So kommt im vorletzten Satz z.B. *als* im Umfeld von *Kollokation* vor, *sprachlicher* im Umfeld von *Zeichen* etc. Wichtig ist, dass dieses gemeinsame Vorkommen nicht zufällig ist. Nun kann man mit Recht behaupten, dass die Wahl eines Wortes in einem durchdachten Text niemals zufällig ist. Wir müssen es also etwas anders formulieren. Wir sprechen von einer Kollokation, wenn ein lexikalisches Zeichen ein anderes lexikalisches Zeichen als Kotext bestimmt, meist unter Ausschluss anderer, bedeutungsähnlicher Zeichen. Der Charakter dieser Auswahl wird deutlich, wenn wir einige in etwa gleichbedeutende Wortverbindungen in verschiedenen Sprachen betrachten. In Tabelle 10 haben wir einige Paare zusammengestellt.

Man sieht an den Daten in Tabelle 26, dass

- die Auswahl eines Wortes durch ein anderes arbiträr und zugleich in einer Einzelsprache konventionalisiert ist, es sich also bei Kollokationen um komplexe sprachliche Zeichen handelt;
- die Auswahl eines Wortes durch ein anderes sich nicht regelhaft beschreiben lässt. Man *putzt* sich die Zähne und *wäscht* sich die Haare oder Hände, man ist mit etwas *hoch* zufrieden oder über etwas *stark* enttäuscht oder gar von etwas *voll* genervt.

[79] Vgl. Eisenberg (2013). Die Seitenzahlen in Klammern beziehen sich auf diese Arbeit.

Diese Wortverbindungen müssen deshalb als Ganzes gelernt bzw. im Wörterbuch gesucht werden.

Sprache 1	Sprache2	Wörtliche Übersetzung
Schlange stehen	spa: hacer cola	Schlange machen
sich die Zähne putzen	fra: se laver les dents	sich die Zähne waschen
den Tisch decken	eng: lay the table	den Tisch legen
dichtes Haar	eng: thick hair	dickes Haar
harte Währung	fra: devises fortes	starke Währung

Tabelle 26: Kollokationen in verschiedenen Sprachen

Als Kollokation im weiteren Sinn hat man im Umfeld des Kontextualismus jedes gemeinsame Vorkommen zweier Wörter im gleichen Kotext bezeichnet[80]. Dieser sehr weite Begriff wird bereits im Umfeld des Kontextualismus weiter eingegrenzt, zunächst auf die Wortpaare, die üblicherweise miteinander vorkommen[81]. Sidney Greenbaum berücksichtigt zudem die syntaktischen Relationen zwischen den miteinander vorkommenden Wörtern[82]. So könnten die Beziehungen zwischen den miteinander vorkommenden Wörtern der Wortklassen Nomen und Adjektiv oder Nomen und Verb gezielt untersucht werden. Die Verbindung von *Ais* und *Kollokation* aus unserem obigen Beispiel würde sich dagegen nicht als Kollokation qualifizieren.

Franz Josef Hausmann schließlich führt den Unterschied zwischen Basis und Kollokator ein. Zwischen diesen beiden Elementen besteht eine gerichtete Beziehung; die Basis bestimmt den Kollokator. Welche Konsequenzen für die Lexikographie das hat, wollen wir an dem Beispiel der Kollokation *schütteres Haar* erläutern. Wenn ein Sprecher oder Schreiber einen Text produzieren möchte, dann ist ihm daran gelegen zu erfahren, welche Prädikate dem Gegenstand *Haar* sprachlich zugeschrieben werden können (z.B. *lang, kurz, blond, rot, braun, graumeliert, strähnig, voll, dicht, schütter*). Dieser potenzielle Benutzer eines Wörterbuchs wird bei der Basis (*Haar*) nachschlagen, um Unsicherheiten bei der Wortwahl zu klären. Hausmann geht es in erster Linie um die Verbesserung der lexikographischen Praxis, die in Einklang zu bringen sei mit den unterschiedlichen Nachschlagebedürfnissen von Benutzern, die einen Text verstehen,

---

80 „[...] innerhalb der britischen Schule des Kontextualismus [...] wurde unter *Kollokation* das faktische Miteinandervorkommen zweier oder mehrerer beliebiger Wörter und/oder lexikalischer Einheiten [...] verstanden [...]. Der Terminus *Kollokation* war in der Theorie des Kontextualismus an keinerlei normative Bewertung hinsichtlich Korrektheit oder Grammatikalität der untersuchten Wortverbindungen gekoppelt.", vgl. Lehr (1996), S. 2.

81 „By collocation is meant the *habitual* association of a word in a language with other particular words in sentences.", vgl. Robins (1964), zit. nach Lehr (1996), S. 5.

82 „A more valuable, if more modest, contribution might be made to the study of collocations if a relatively homogenous class of items were selected and an investigation undertaken of the collocation of each item in the class with other items that are related syntactically in a given way.", vgl. Greenbaum (1970), S. 13.

und Nutzern, die einen Text erstellen wollen[83]. Wir teilen Hausmanns Meinung, dass es sinnvoll ist, dem Begriff *Kollokation* ein schärferes Profil zu geben. Für sprachtechnologische Zwecke aber mag es genügen, die Wortpaare zu finden, die häufiger als erwartbar miteinander vorkommen. Um beiden Phänomenen gerecht zu werden, wollen wir hier zwischen *Kookkurrenz* und *Kollokation* (im engeren Sinn) unterscheiden.

- Als *Kookkurrenz* soll das gemeinsame Vorkommen zweier Wörter in einem gemeinsamen Kotext betrachtet werden. Die Länge des betrachteten Kotextes kann als Textfenster einer bestimmten Länge festgelegt werden. Im Allgemeinen wird vom einzelnen Beleg abstrahiert und das gemeinsame Vorkommen zweier Wörter in vielen Kotexten betrachtet. Es kann zudem die Reihenfolge des Auftretens beider Wörter in den Belegen als unterscheidendes Kriterium zweier Kookkurrenzen festgelegt werden[84]. Ferner kann festgelegt werden, dass die Wörter einer Kookkurrenz häufiger (im gegebenen Textfenster) miteinander vorkommen, als dies der Fall wäre, wenn die Wörter zufällig verteilt wären. Man spricht in diesem Fall von einem *signifikanten* Kovorkommen beider Wörter und verwendet statistische Assoziationsmaße, um dies zu messen[85].
- Eine *Kollokation* muss natürlich den oben genannten Kriterien genügen, darüber hinaus aber auch eine innere Struktur, in Form einer Hierarchie zwischen Kollokationsbasis und Kollokator aufweisen. Darüber hinaus müssen die Glieder einer Kollokation in einer syntaktischen Beziehung zueinander stehen, z.B. als Kopf einer Verbalphrase und Kopf einer gleich- oder untergeordneten Nominalphrase, oder als Kopf einer Nominalphrase und Kopf einer untergeordneten Adjektivphrase[86].

Es ist offensichtlich, dass Korpora für das Aufspüren von Kookkurrenzen und Kollokationen von großem Nutzen, wenn nicht gar unverzichtbar sind. Je größer das Korpus, desto mehr Belege für ein beliebiges Wortpaar wird man darin finden. Dies macht die darauf basierenden Statistiken zuverlässiger. Im einfachsten Fall, dem der Kookkurrenz, reicht es, das Korpus in eine Menge von Textfenstern aufzuteilen und zu ermitteln: a) in wie vielen Fenstern $Wort_1$ und $Wort_2$ gemeinsam vorkommen, b) in wie vielen Fenstern nur $Wort_1$ vorkommt, c) in wie vielen Fenstern nur $Wort_2$ vorkommt und d) in wie vielen Fenstern weder $Wort_1$ noch $Wort_2$ vorkommen. Die meisten Assoziationsmaße setzen diese vier Werte bzw. ihre Summen miteinander in Beziehung. Das Ergebnis der Anwendung eines Assoziationsmaßes auf ein Wortpaar ist eine Kennziffer, durch die dieses Wortpaar mit anderen Wortpaaren in Beziehung gesetzt werden kann. Wortpaare mit hohen Kennziffern sind signifikante Kookkurrenzen und damit gute Kandidaten für Kollokationen. Die anderen Bedingungen für eine Kollokation müssen allerdings auch gegeben sein. Um dies zu prüfen, braucht man ein Korpus, bei dem zumindest die Wortarten annotiert sind, oder eine Belegsammlung.

---

[83] Zu dieser Position vgl. vor allem Hausmann (1985) und Hausmann (2004).

[84] Die Wortfolge *doch eben* bedeutet eben doch etwas anderes als die Wortfolge *eben doch*.

[85] Eine Übersicht über statistische Assoziationsmaße geben Lemnitzer (1997), Kapitel 4, und Evert (2004).

[86] Einige Beispiele für diese Beziehungen befinden sich in Tabelle 26.

Elisabeth Breidt wendet ein solches Verfahren auf ein wortartenannotiertes Korpus an, um Nomen-Verb-Kollokationen zu ermitteln[87]. Lothar Lemnitzer[88] experimentiert mit verschiedenen Assoziationsmaßen und arbeitet ebenfalls mit einem wortartengetaggten Korpus und exemplifiziert dessen Nutzen am Beispiel der Kollokanten des lexikalischen Zeichens *hart*[89]. Joachim Wermter und Udo Hahn extrahieren Kollokationen zwischen Präpositionalphrasen und Verben aus einem großen, ebenfalls wortartengetaggten Korpus[90]. Von hoher praktischer Relevanz sind schließlich auch die Arbeiten am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart. Stellvertretend sei hier auf die Arbeit von Heike Zinsmeister und Ulrich Heid hingewiesen[91]. Die Autoren extrahieren aus einem getaggten und partiell geparsten Zeitungskorpus Kombinationen von Verb, Nomen und modifizierendem Adjektiv, trennen die relevanten von den irrelevanten Kombinationen und klassifizieren die relevanten Tripel halbautomatisch in sechs Klassen, die das Spektrum von der idiomatischen Wendung (z.B. *offene Türen einrennen*) bis zur gänzlich freien Fügung (z.B. *konkrete Zahlen nennen*) abdecken. Die Relevanz dieser Arbeit für die praktische Lexikographie ist offensichtlich. Die Autoren diskutieren auch die Grenzen und Probleme ihres Ansatzes. So gibt es zur Zeit kein Verfahren, das auf der Basis der Unterschiede der sechs Klassen eine vollständige und vollkommene Klassifizierung erreichen kann[92].

Beim Digitalen Wörterbuch der deutschen Sprache wird für die Ermittlung von Kollokationen für die zu bearbeitenden Stichwörter ein Wortprofil verwendet[93]. Diese Software ist in vielerlei Hinsicht mit der ‚Word Sketch Engine' von Adam Kilgarriff vergleichbar[94] und hat die folgenden Eigenschaften: a) es werden zu Wörtern, die häufig genug in den zugrunde liegenden Korpora vorkommen, die Kookkurrenzen ermittelt; b) da die zugrunde liegenden Korpora syntaktisch geparst und annotiert sind, sind die Kookkurrenzen nach den syntaktischen Relationen, in denen die Wortpaare stehen, geordnet; c) der zugeordnete Salienzwert bezieht sich auf ein Wortpaar in der jeweiligen syntaktischen Relation (wird also lokal und nicht global bestimmt); d) die Ergebnisse können als Tabelle, geordnet nach Salienz des Paares, oder als Wortwolke angezeigt werden; e) es ist eine vergleichende Analyse für zwei Kollokationsbasen möglich – in dem Fall werden einerseits die für jeweils eine Kollokationsbasis typischen Kookkurrenzen, andererseits die für die beiden Basen gleich typischen Kookkurrenzen angezeigt.

Die korpusbasierte Untersuchung von festen Redewendungen, *Phraseme* genannt, steht hinter der Untersuchung von Kollokationen bisher deutlich zurück. Eine Ausnahme bildet eine Arbeitsgruppe um Christiane Fellbaum an der Berlin-Brandenburgischen Akademie der Wissenschaften. Diese Gruppe hat es sich zum Ziel gesetzt, systematisch,

---

87 Vgl. Breidt (1993).

88 Vgl. Lemnitzer (1997), Kap. 4.

89 Vgl. Lemnitzer (1997), Kap. 5. Das Hauptziel dieser Arbeit ist es, korpusgestützt Mehrwortlexeme zu ermitteln, Kollokationen sind dort nur ein Aspekt unter mehreren.

90 Vgl. Wermter und Hahn (2004). Zwei ihrer Beispiele sind *unter Druck geraten* und *in den Griff bekommen*.

91 Vgl. Zinsmeister und Heid (2003).

92 Einige Beispiele für alle sechs Klassen werden in Zinsmeister und Heid (2003), Tabelle 5, präsentiert.

93 Vgl. Geyken (2011).

94 Vgl. Kilgarriff et al. (2004).

möglichst vollständig und mit synchroner und diachroner Perspektive die Gruppe der aus einer Verbalphrase und einer untergeordneten Nominalphrase bestehenden Phraseme zu untersuchen[95]. Phraseme zeichnen sich dadurch aus, dass

- sie nach der Grammatik der entsprechenden Sprache nicht immer wohlgeformt sind (z.B. *ganz Ohr sein*);
- sie semantisch intransparent sind, die einzelnen Bestandteile also nicht die Bedeutung haben, die sie in freier Verwendung haben (z.B. *die Katze aus dem Sack lassen*);
- sie nur begrenzt modifizierbar sind (vgl. *einen Kater haben, einen furchtbaren Kater haben, einen grau gescheckten Kater haben*, im letzten Fall geht die idiomatische Lesart – unter den Folgen überhöhten Alkoholgenusses leiden – verloren)[96].

Die von Fellbaum und ihrem Team untersuchten verbalen Phraseme zeichnen sich dadurch aus, dass sie oft komplexe Sachverhalte benennen und deshalb nicht einfach in die semantischen Strukturen des Lexikons einer Sprache eingefügt werden können[97].

In einer detaillierten Arbeit untersuchen sie die Funktion der hochgradig unspezifischen Pronomen *etwas* und *ein(en)* als Ergänzungen verbaler Phraseme[98]. In manchen Fällen haben diese Ergänzungen Argumentstatus und referieren auf etwas, wenn auch sehr Unspezifisches (z.B. *etwas auf der hohen Kante haben*). In anderen Fällen hat *etwas* keinen Argumentstatus und referiert nicht (z.B. *jemandem etwas husten*). Die Autoren vermuten, dass der „Platzhalter" hier grammatische Funktionen übernimmt. Zum einen ermöglicht er die Einführung eines indirekten Objekts (das die Existenz eines direkten Objekts voraussetzt; *etwas* füllt diesen Platz aus). Zum anderen erzwingt *etwas* die Interpretation des Verbs und damit des gesamten Phrasems als zeitlich eingegrenztes Ereignis. Zwischen diesen beiden Verwendungen von *etwas* gibt es, wie die Autoren zeigen, etliche Zwischenstufen. Ähnliche Befunde werden bei der Analyse von *ein(en)* ermittelt.

Die Arbeit ist vor allem für die lexikographische Praxis relevant. Da die beiden Hauptfunktionen von *etwas* und *ein(en)* die möglichen Modifikationen des Phrasems im Gebrauch beeinflussen, sollten bei der lexikographischen Ansatzform diese beiden Elemente zumindest graphisch unterschieden werden[99].

8.5.4 Partikeln

Eine korpuslinguistisch sehr gute bearbeitete Wortart sind die Partikeln. Wir wollen hier die wichtigsten korpusbezogenen Arbeiten als Beispiele für korpusbasierte linguistische Forschung im Bereich der Wortarten vorstellen.

---

[95] Vgl. Fellbaum (2002), Abschnitt 6.

[96] Für eine detaillierte Analyse vgl. Keil (1997).

[97] *einen zwitschern* ist eben mehr als eine bestimmte Art zu trinken, das Phrasem evoziert eine ganze Szene, bei der das Trinken alkoholischer Getränke eine Rolle spielt. Dieses „mehr" ist es, was die Forscher vor allem interessiert, vgl. Fellbaum (2002), Abschnitt 3.

[98] Vgl. Fellbaum et al. (2004).

[99] Vgl. Fellbaum et al. (2004), Abschnitt 5.

Es herrscht weitgehend Uneinigkeit darüber, welche Wörter zu den Partikeln zählen und in welche Unterklassen diese Wortklasse zerfällt. Die Duden Grammatik[100] subsumiert die Adverbien, Präpositionen und Konjunktionen unter die Partikeln und wählt damit eine sehr weite Definition, die die meisten nicht flektierenden Wörter umfasst[101]. In einem engeren Sinn verwendet etwa Helbig diesen Begriff[102]. Er bezeichnet mit *Partikel* „solche morphologisch unflektierbaren Wörter, die über keine solchen syntaktischen Funktionen verfügen, wie sie den Wörtern anderer unflektierter Wortklassen zukommen" (S. 20). Eine noch engere Definition fasst lediglich die Modalpartikeln in diese Kategorie (S. 21). Helbig unterscheidet die folgenden Subklassen von Partikeln:

- Abtönungs- oder Modalpartikeln (z.B. *auch, bloß, denn*);
- Gradpartikeln (z.B. *auch, gerade, sogar*);
- Steigerungspartikeln (z.B. *außerordentlich, etwas, ganz*);
- Temporalpartikeln (z.B. *erst, noch, schon*);
- Antwort- oder Satzpartikeln (z.B. *ja, doch, eben*);
- Vergleichspartikeln (z.B. *wie, als*);
- Interjektionspartikeln (z.B. *au, oh je*);
- Negationspartikeln (z.B. *kein, nicht*);
- Infinitivpartikel *zu*.

Wie kaum eine andere Wortart beziehen die Partikeln ihre Bedeutung durch ihren Kontext. Die Partikeln tragen nichts zur propositionalen Bedeutung einer Äußerung bei, wie man an dem folgenden Beispiel sieht:

(14) Was macht Peter jetzt (*eigentlich*)?

Auch ohne die Modalpartikel *eigentlich* ist der Satz als Frage über Peters momentane, z.B. berufliche, Aktivitäten verständlich. Die Partikel erfüllt hier die Funktion, die Frage als Eröffnung eines neuen Themas zu markieren, z.B. in einem Dialog wie dem folgenden:

(15) A: Stell dir vor, da steht: 100 km Stau! – B: Na, da wirds wieder gekracht haben. – A: Sag mal, wie hoch ist man eigentlich versichert, wenns mal so richtig kracht?[103]

Sie hat hier also gesprächssteuernde Funktion und sichert außerdem die Kohärenz, die bei einem abrupten Themenwechsel sonst gefährdet wäre.

Modalpartikeln können außerdem dazu verwendet werden, um die Haltung des Sprechers zum Beispiel zum Wissen oder den Haltungen der Gesprächspartner zu signalisieren:

(16) Und sie bewegt sich *doch*.

(17) Das kann schon mal etwas wackeln. Sie bewegt sich *ja*.

---

[100] Vgl. Fabricius-Hansen et al. (2009).

[101] Dies stimmt nur so ungefähr, da den Interjektionen ein eigenes Kapitel gewidmet ist.

[102] Vgl. Helbig (1994).

[103] Dieses leicht modifizierte Beispiel entstammt Thurmair (1989), S. 176.

In Beispiel (16) signalisiert der Sprecher seine Annahme, dass die Gesprächspartner seine Behauptung (bisher) nicht teilen, in Beispiel (17) hingegen wird signalisiert, dass das Behauptete auch den Gesprächspartnern bekannt ist. Die Redundanz der Äußerung wird dadurch abgemildert.

Die Analyse und Beschreibung von Partikeln ist eine besondere Herausforderung für die theoretische Linguistik, die Lexikographie, die maschinelle Sprachverarbeitung und für die Fremdsprachvermittlung[104].

Viele Partikeln stellen die theoretische Linguistik und Lexikographie vor schwierige Aufgaben: Da sie nichts zur Proposition einer Äußerung, in der sie auftauchen, beitragen, müssen die kontextreferentiellen Funktionen dieser sprachlichen Einheiten bestimmt und beschrieben werden. Dies muss in so allgemeiner Weise geschehen, dass möglichst alle Verwendungsweisen bzw. Verwendungssituationen mit dieser Beschreibung abgedeckt werden. Die Gefahr einer solchen generischen Beschreibung ist, dass sie zu allgemein und damit wertlos wird. Will man andererseits das Spezifische der Verwendungskontexte aller Modalpartikeln erfassen, läuft man Gefahr, das Gemeinsame aller Verwendungsinstanzen in den Details zu verlieren.

Das Problem der Analyse von Partikeln stellt sich in verschärftem Maße bei der maschinellen Analyse natürlicher Sprache. Die spezifische „Bedeutung" bzw. ihre pragmatische Funktion kann nur erfasst werden, wenn Wissen über den Kontext der jeweiligen Äußerung vorhanden ist. Selbst wenn dieses Kontextwissen nicht in ein Computerlexikon gehört, so doch zumindest eine lexikalische Beschreibung, die Angaben zu den möglichen Vorkommenskontexten umfasst.

Viele lexikalische Elemente der Partikelklasse gehören mehreren Unterklassen an und einige darüber hinaus auch anderen Wortklassen[105]. Die Verwendung dieser lexikalischen Elemente muss einerseits von einander abgegrenzt, andererseits miteinander in Beziehung gesetzt werden. Einige Typen von Partikeln, z.B. die Modalpartikeln, treten zudem in zahlreichen Kombinationen auf. Maria Thurmair[106] listet weit über 100 Kombinationen auf, von denen viele aber nur eingeschränkt akzeptabel seien.

All diese Aspekte von Partikeln machen diese zu einem besonders guten Gegenstand für korpuslinguistische Untersuchungen.

Umfangreiche linguistische und lexikologische Studien zu den Partikeln von Harald Weydt erschienen bereits 1979 und Anfang der 1980er Jahre[107]. Gerhard Helbig[108] widmet den Partikeln ein eigenes Wörterbuch. Darüber hinaus ist die Frage der angemessenen Übersetzung der Abtönungspartikeln in eine andere Sprache ein wichtiges Problem. König, Stark und Requardt füllen eine Lücke mit ihrem deutsch-englischen Spezialwörterbuch zum Wortschatzbereich der Adverbien und Partikeln[109].

---

[104] Wir gehen im Abschnitt zum Fremdspracherwerb und -vermittlung u.a. auf eine Arbeit zur didaktischen Vermittlung des Gebrauchs von Modalpartikeln ein.

[105] So kann z.B. *doch* als Antwortpartikel und als Modalpartikel entsprechend der Klassifikation von Helbig und als Konjunktion verwendet werden.

[106] Vgl. Thurmair (1989).

[107] Vgl. Weydt (1979), Weydt (1983) u.a.

[108] Helbig (1994).

[109] Vgl. König et al. (1990).

Mittlerweile sind zahlreiche Monographien und detaillierte Arbeiten auch zu einzelnen Partikeln oder zu Partikelgruppen erschienen[110]. Von besonderem Interesse sind dabei die Bedeutungs- oder Funktionskontraste nah verwandter Partikeln[111]. Es wurde aber zu Recht kritisiert, dass linguistische Arbeiten zu den Partikeln allzu oft auf erfundene oder konstruierte Beispiele aufbauen[112]. Diese mögen als Testmaterial zur Ermittlung von Akzeptabilitätsurteilen oder zur Ermittlung von Kontrasten geeignet sein, erscheinen aber in vielen Fällen unnatürlich und können nicht das wiedergeben, was in authentischen Gesprächen geschieht[113]. Die Situation hat sich in den letzten Jahren gebessert, was auch durch die bessere Verfügbarkeit von Korpora geschriebener, vor allem aber auch gesprochener Sprache bedingt ist.

Thurmair, die in einer Monographie die Kombinierbarkeit von Modalpartikeln untersucht[114], erwähnt einen Vorschlag von Collins aus dem Jahre 1938:

> It would be an alluring task to pick out in German a certain number of simple particles, combine them in pairs or triplets or even larger groups, and try to discover which groups are the most commonly used, which have the most characteristic functions, and which cannot be combined with which, or at least not in a particular order. (Zit. nach Thurmair 1989, S. 203)

Tatsächlich ist durch die Existenz sehr großer Korpora und der entsprechenden Werkzeuge zu ihrer Analyse nun die Möglichkeit gegeben, zumindest eine der von Collins gestellten Fragen zu beantworten, nämlich die nach der Vorkommenshäufigkeit, Reihenfolge und Bindungsstärke einzelner Partikelkombinationen.

Collins schneidet außerdem wichtige Fragen an, mit deren Beantwortung erst begonnen wurde:

- Sind die Restriktionen, denen die Kombinierbarkeit von Abtönungspartikeln unterliegt, systematisch zu beschreiben? Dies betrifft sowohl die Möglichkeit des Kovorkommens zweier Abtönungspartikeln als auch die Reihenfolge ihres Auftretens. Die allgemeinste Beschränkung des Kovorkommens ist dadurch gegeben, dass zwei Partikeln, deren Modus inkompatibel ist, nicht zusammen auftreten.[115]
- Wenn zwei Abtönungspartikeln miteinander in einem Satz vorkommen können, ist ihre Abfolge durch Prinzipien beschreibbar, die sich aus ihren Merkmalen ergeben? Helbig bildet zwar Distributionsklassen für eine Reihe der Partikeln, um deren Reihenfolgebeziehung bei der Kettenbildung zu erfassen, die Belege für seine Hypothesen sind allerdings wenig überzeugende Eigenkonstruktionen.[116]

Thurmair verwendet eine Menge von semantischen Merkmalen, nach denen die einzelnen Partikeln im ersten Teil ihrer Arbeit klassifiziert werden. Mit diesen Mitteln sol-

---

110 Einen guten Überblick gibt das Literaturverzeichnis in Moellering (2004).

111 Zum Beispiel die der Gradpartikeln *auch* und *noch*, die im Mittelpunkt der Monographie von Ulrike Nederstigt (2003) stehen.

112 Unter anderem von Ulrike Nederstigt, vgl. Nederstigt (2003).

113 Vgl. Nederstigt (2003), S. 12.

114 Vgl. Thurmair (1989).

115 Vgl. Helbig (1994), S. 76 und Thurmair (1989), S. 204f.

116 Vgl. Helbig (1994), S. 75f.

len Selektionsrestriktionen und Kombinationspräferenzen beschrieben werden. Diese semantischen Merkmale gehen in die Partikelkombinationen ein.

Die Einzelbedeutungen der Partikeln werden nach Auffassung der Autorin zur Gesamtbedeutung der Partikelkombinationen addiert:

> Es soll hier davon ausgegangen werden, daß eine Kombination der Partikel A mit der Partikel B eine Addition ihrer Bedeutung und damit ihrer Merkmale bedeutet; d.h. also, daß sich die Kombinationen in ihre Einzelteile zerlegen lassen. (Thurmair 1989, S. 205)

Thurmair führt dieses Programm in ihrer Monographie dadurch aus, dass sie die Bedeutung und Funktion der einzelnen Modalpartikeln[117] und im Anschluss daran die akzeptablen, bedingt akzeptablen und inakzeptablen Kombinationen beschreibt[118]. Sie beschließt ihre Arbeit mit einer Synopse der Partikelkombinationen (S. 278), mit einer Übersicht über die Distribution der Kombinationen über verschiedene Satztypen (S. 282, Tabelle 13) und einer Übersicht über Stellungsregeln für einzelne Modalpartikeln (S. 285–289). Ob das Postulat der additiven Bedeutung von Partikelkombinationen durchzuhalten ist, bleibt unklar. Hier setzt die Kritik von Lothar Lemnitzer[119] an, der davon ausgeht, dass Partikelkombinationen als komplexe Lexeme nicht transparent und analysierbar sind. Lemnitzer präsentiert als Fallstudie die Kombinationen mit der Modalpartikel *denn*[120] und untersucht, ähnlich wie Thurmair, systematisch alle Kombinationen, konzentriert sich aber bei den Einzeldarstellungen auf die häufigsten. Er trifft allerdings keine Aussage zur psychologischen Plausibilität seiner Vermutung, dass es sich hier um Mehrwortlexeme handelt. Im Zentrum seiner Arbeit steht vor allem die Analyse und (computer-)lexikographische Erfassung dieser komplexen sprachlichen Einheiten.

Die Aneignung des korrekten Gebrauchs von Partikeln im Zuge des Erstspracherwerbs verfolgt Ulrike Nederstigt am Beispiel der Gradpartikeln *auch* und *noch*[121]. Der Erwerb des komplexen sprachlichen Wissens, das für die korrekte Verwendung der Partikeln notwendig ist, demonstriert die Autorin am Gebrauch dieser Partikeln durch erwachsene Sprecher. Verschiedene linguistische Versuche, die phonologischen, syntaktischen, semantischen und diskursiven Aspekte des Gebrauchs dieser Partikeln zu beschreiben und zu erklären, stellt die Autorin auf den Prüfstand. Sie verwendet hierfür mehrere Korpora gesprochener Sprache, von denen einige den Sprachgebrauch Erwachsener, andere den Sprachgebrauch eines Mädchens in der Phase zwischen dem zweiten und vierten Lebensjahr wiedergeben[122]. Die Verwendung von Korpora, vor allem solcher der gesprochenen Sprache, erlaubt ihr, gemessen an den bisherigen linguistischen Arbeiten zum Thema, einen neuen Blick auf die von ihr beschriebenen Partikeln[123]. Die wichtigsten Erkenntnisse sind:

---

117 Abschnitt 2, S. 94–202.

118 Abschnitt 3, S. 203–284.

119 Vgl. vor allem Lemnitzer (2001).

120 Vgl. Lemnitzer (1997).

121 Vgl. Nederstigt (2003). Die Autorin spricht von „focus particles", dies entspricht aber der von Helbig definierten Klasse der Gradpartikeln.

122 Zu den verwendeten Korpora vgl. Kapitel 3.2, S. 80–83 (Korpora der Erwachsenensprache) und Kapitel 7.2, S. 211 (Korpus der Kindersprache) in Nederstigt (2003).

123 Kapitel 6 ist überschrieben mit „A fresh look at focus particles".

- Die Gradpartikel *noch* hat mehr Bedeutungen als gemeinhin angenommen, es werden insgesamt neun Bedeutungen unterschieden:
  - additiv (*Nimm dir noch einen Nachtisch, bitte!*);
  - additiv, weiteres Element einer Menge (*Davor habe ich noch einen Termin.*);
  - additiv, vor einer Wende (*Damals haben wir noch ohne fließend Wasser gewohnt.*);
  - temporal (*Der Platz ist noch frei.*);
  - temporal, perfektiv (*Wir können uns gern noch in diesem Monat treffen.*);
  - temporal, mit abnehmender Menge von Objekten (*und jetzt brauche ich noch eine Siebenerleiste.*);
  - mit *einmal*, repetitiv (*Dann sind wir noch einmal Karussell gefahren.*);
  - mit *einmal*, restitutiv (*Knöpf die Jacke am besten noch einmal auf.*);
  - mit *einmal*, additiv (*Wir sollten uns dann noch einmal Zeit nehmen, wenn das heute nicht klappt.*)[124].

  Der Kontrast zwischen diesen Bedeutungen korrespondiert mit Unterschieden in den phonologischen und syntaktischen Merkmalen der Partikel bzw. mit Unterschieden in den Verwendungskontexten[125].
- Bei der Gradpartikel *auch* unterscheidet die Autorin zwischen einer betonten Variante und einer unbetonten Variante. Die betonte Variante weist Merkmale einer Antwortpartikel auf. Ihr Gebrauch signalisiert, dass einer vorhergehenden positiven (oder negativen) Antwort eine weitere positive (oder negative) Antwort hinzugefügt wird, wie in dem folgenden Beispiel:

  (18) Mitte der Woche habe ich AUCH nicht so gerne ...

  Der Kontext dieser Äußerung ist die telefonische Suche nach einem Termin für ein Treffen. Der Sprecher hat bereits vorher einige Terminvorschläge negativ beschieden. Das Beispiel wird hier als eine weitere Ablehnung interpretiert[126]. Die unbetonte Variante weist die typischen Merkmale einer Gradpartikel auf, besonders eine größere Variabilität in der Wortstellung, wie die beiden folgenden Beispiele zeigen[127].

  (19) Wir können sonst auch Freitag oder Samstag nehmen.

  (20) Auch Freitag oder Samstag können wir sonst nehmen.

- Wenn man die betonte und unbetonte Variante der Fokuspartikel *auch* als zwei verschiedene lexikalische Elemente betrachtet, dann wird eine homogene Beschreibung der Funktionen der unbetonten Variante und der Fokuspartikel *noch* möglich. Es bleiben semantische Unterschiede zwischen beiden Partikeln, die aber ihrer Subsumierung in eine Klasse nicht entgegenstehen.

---

[124] Vgl. ebd., S. 100–106. Die hier gewählten Beispiele sind erfunden, Nederstigt präsentiert authentische, aber auch etwas komplexere Beispiele.

[125] Vgl. ebd., Kapitel 4.2.2, S. 167–171.

[126] Vgl. ebd., S. 190f. Die Großschreibung der Partikel in unserer Wiedergabe des Beispiels signalisiert, dass sie betont ist.

[127] In Anlehnung an Nederstigt, weitere Beispiele in ihrer Arbeit auf S. 200.

Die Untersuchung von Spracherwerbsdaten in Hinblick auf die beiden Partikeln – und der Partikel *auch* in beiden Betonungsvarianten – zeigt, dass die Unterscheidung der beiden Varianten von *auch* kognitiv plausibel ist: Die betonte Variante von *auch* wird früher erworben als die unbetonte Variante, und letztere wiederum in etwa zur gleichen Zeit wie die Partikel *noch*[128]. Nederstigt stützt ihre Erkenntnisse auf Langzeitaufzeichnungen der Sprachentwicklung eines Kindes namens *Caroline*. Caroline beginnt mit einem Jahr und neun Monaten, *AUCH* zu verwenden, und mit gut zwei Jahren, ungefähr zur gleichen Zeit, *auch* und *noch*.

Durch die gründliche qualitative Analyse von Korpusbelegen, die aus mehreren Korpora gesprochener Sprache stammen, gelingt es der Autorin, hinsichtlich einer der beiden untersuchten Partikeln (*auch*) eine Unterscheidung zwischen zwei Varianten zu treffen. Der Schnitt, den sie macht, erlaubt es, eine Variante dieser Partikel konsistent in das System der Fokuspartikeln zu integrieren.

8.6 Computerlinguistik

Die Computerlinguistik ist ein Bereich, in dem Korpora eine wichtige Rolle spielen[129]. Zunächst dienen sie als Datenquelle für das empirische Arbeiten. Der Computerlinguist sichtet Korpusdaten, um seine Hypothesen, Modelle oder Programme an authentischem Material zu entwickeln und zu prüfen. In diesem Vorgehen unterscheidet er sich nicht von anderen Linguisten.

Der Unterschied besteht darin, dass der Computerlinguist Korpora auch in großem Maßstab zum Entwickeln und Prüfen seiner Programme nutzen kann. Was ist damit gemeint?

Bei der Entwicklung von Programmen nutzt er die Frequenzinformationen, die in einem Korpus stecken, z.B. beim *Training* von statistischen Programmen[130]. Diese Programme beinhalten Regeln, deren Anwendungen über so genannte Gewichte gesteuert werden. Eine Regel mit höherem Gewicht wird bevorzugt angewendet. Die Werte für die Gewichte werden aus Korpora abgeleitet, indem man die Wahrscheinlichkeiten für die Regeln anhand eines Korpus ermittelt (in der Computerlinguistik sagt man, das Programm *lernt* die Wahrscheinlichkeiten beim *Training*). Stark vereinfacht zählt das Programm dabei, wie oft eine Regel bei der Analyse des Korpus angewendet wird[131].

Ein Beispiel für das Lernen aus Korpora ist die *Grammatikinduktion*. Aus den Annotationsstrukturen eines Korpus werden Frequenzen für Grammatikregeln abgelesen. Im Extremfall leitet man sogar die Grammatikregeln selbst aus dem Korpus ab (Anette Frank[132] erzeugte z.B. eine lexikalisierte *Tree Adjoining Grammar* auf der Basis des NEGRA-Korpus).

---

[128] Vgl. ebd., Abschnitt 9.4.1, vor allem Abbildung 9.7 auf S. 340.

[129] Ein guter Überblick über das Verhältnis von Computerlinguistik und Korpuslinguistik liegt mit Dipper (2008) vor.

[130] Im vierten Kapitel hatten wir Ihnen im Exkurs zum Part-of-Speech Tagging z.B. das Training des Brill-Taggers vorgestellt.

[131] Zwei empfehlenswerte englischsprachige Einführungen zur statistischen Sprachverarbeitung sind Jurafsky und Martin (2000) und Manning und Schütze (1999).

[132] Vgl. Frank (2001).

Das Training kann auch unter indirekter Nutzung eines Korpus stattfinden. Manchmal werden zuerst Daten aus einem Korpus extrahiert und zum Beispiel in einer Datenbank gesammelt. Die im Abschnitt zur Syntax beschriebene Arbeit von Timm Lichte ist ein Beispiel dafür. Lichte listet Kookkurrenzen von Wörtern und Lizenzierern für Negative Polaritätselemente auf, um mit Hilfe eines statistischen Programms Kandidaten für Negative Polaritätselemente zu bestimmen.

Sabine Schulte im Walde[133] zeigt, wie man mit computerlinguistischen Methoden die Verbklassen von Levin (1993) auf deutschen Daten nachvollziehen kann. Sie trainiert zunächst eine Grammatik auf dem Huge German Corpus, um Frequenzinformationen über Verben, deren Argumentrahmen und die aufgetretenen nominalen Realisierungen der Argumente zu erfassen. In einem zweiten Schritt entwickelt sie ein Programm, das aus diesen Informationen Klassen von Verben bilden kann (das Programm *clustert* die Verben in Gruppen), z.B.[134]:

(21) *Verben, die sich auf eine Basis beziehen:*
basieren, beruhen, resultieren, stammen

(22) *Verben der Maßänderung:*
reduzieren, senken, steigern, verbessern, vergrößern, verkleinern, verringern, verschärfen, verstärken, verändern (...)

Eine weitere Verwendungsweise von Korpora in der Computerlinguistik ist das Testen von Programmen, anders ausgedrückt die *Evaluierung*. Hierzu benötigt man ein linguistisch annotiertes Korpus (den *Goldstandard*), das idealerweise mit den Strukturen annotiert ist, die das Programm erzeugen soll. Der Idealfall ist allerdings nicht immer gegeben, da – wie Sie ja inzwischen wissen – Annotation sehr aufwändig und kostenintensiv ist. Man muss manchmal Kompromisse eingehen und z.B. die Ausgabe des eigenen Programms auf das vorgegebene Format des Testkorpus abbilden. Letzteres hat den einen Vorteil, dass man auf diese Art verschiedene Programme unmittelbar anhand desselben Testkorpus vergleichen kann. Wenn man testet, muss man sich klar machen, dass auch das Testkorpus Fehler enthalten kann. Es bietet sich daher an, als obere Grenze bei einer Evaluierung nicht 100% Übereinstimmung zu verlangen, sondern sich an der Übereinstimmung der Annotatoren des Goldstandards zu orientieren (am *Inter-Annotator Agreement*).

8.7 Fremdspracherwerb und -vermittlung

Im sechsten Kapitel haben wir die Dichotomie von Korpora im Fremdspracherwerb und -vermittlung schon erwähnt: Sie umfassen sowohl muttersprachliche Korpora, die als Datenressource im Unterricht eingesetzt werden können, als auch Korpora, die den Fremdspracherwerb dokumentieren, also Sprache von Nichtmuttersprachlern enthalten.

---

[133] Vgl. Schulte im Walde (2003).

[134] Wir stellen hier nur korrekte Beispiele vor, um das Ergebnis zu veranschaulichen. Das Programm clustert teilweise auch Verben in eine Gruppe, die keine gemeinsame Bedeutung besitzen.

Joybrato Mukherjee[135] beschreibt in seiner Einführung in die Korpuslinguistik ausführlich, wie Korpora für den Englischunterricht eingesetzt werden können, sowohl in der Unterrichtsvorbereitung als auch im Unterricht selbst. Sie dienen als Quelle für natürliche Beispiele und geben dem Sprachlerner frühzeitig Kontakt zur natürlichen Sprachverwendung.

Diese Verwendungsweise bietet sich insbesondere auch für die Erstellung von Lehrbüchern an. Dieter Mindt[136] analysiert Lehrbücher für den Englischunterricht, die an deutschen Schulen eingesetzt werden, und stellt fest, dass sie teilweise irreführend dahingehend sind, dass weniger häufig verwendete Formen früher eingeführt werden als die eigentlich gängigen. Dadurch entsteht beim Lernen ein falsches Gewicht. Als Negativbeispiel stellt er das Englische *going to*-Futur vor, das in mehreren Standardlehrbüchern früher eingeführt wird als das viel häufiger verwendete *will*-Futur. Er argumentiert, dass Lehrwerke, die auf der Basis von korpusbestimmten quantitativen Untersuchungen von Wortschatz und Verwendungsweisen erstellt werden, solche Verzerrungen nicht enthalten[137]. Guy Aston[138] nennt diese Verwendung von Korpora in der Lehre, bei der der Sprachlerner keinen direkten Zugang zu den Korpora bekommt, den *Hinter den Kulissen*-Ansatz ('Behind the Scenes Approach'). Er kontrastiert ihn mit dem *Auf der Bühne*-Ansatz ('On Stage Approach'), bei dem der Lerner direkt mit dem Korpus arbeitet. Bei diesem Ansatz kann *Data-Driven Learning* zum Einsatz kommen, d.h. Lerner leiten von den Daten Generalisierungen ab, die sie dann auf die Analyse neuer Daten anwenden. Die Analyse von Sprache wird so direkt mit ihrer natürlichen Verwendung gekoppelt. Technische Voraussetzungen dafür sind ein Korpus, ein Konkordanzwerkzeug und Werkzeuge zur eigenen Datenextraktion. Konkrete Anwendungsszenarien sind das Nachschlagen von Wortverwendungen im Satzkontext für die Textproduktion und -rezeption, das systematische Untersuchen bestimmter Sprachverwendungen oder Grammatikkonstruktionen und das „genüssliche Schmökern" ('serendipitous exploration'). Sogar eine Art enzyklopädischer Verwendung ist möglich, da man durch das Korpus Informationen zu bestimmten Orten oder Personen erhalten kann, sowie über die Kultur der Sprachgemeinschaft, wenn z.B. nach Stereotypen und Vorurteilen geforscht wird[139]. Aston geht auch auf begleitende Effekte des Korpuseinsatzes im Klassenzimmer ein, z.B. den kommunikativen Aspekt bei gemeinsamer Korpusarbeit (Korpusanfrage, Finden von Mustern, Interpretation usw.). Als Zielgruppe für diese Art von Korpuseinsatz im Unterricht empfiehlt er fortgeschrittene (erwachsene) Lerner und Lehrer, da es z.B. schwieriger ist, Konkordanzzeilen zu interpretieren, als ein Lernerwörterbuch zu lesen[140].

Im dänischen *Visual Interactive Syntax Learning*-Projekt (kurz: VISL)[141] kommen linguistisch annotierte Korpora direkt zum Einsatz, wenn auch nicht ganz offen „auf der

---

[135] Vgl. Mukherjee (2002).

[136] Vgl. Mindt (1996).

[137] Bereits seit 1980 werden im Rahmen des *COBUILD*-Projekts – eine Kooperation zwischen einem Verlag und der Universität Birmingham – in einem korpusbasierten Ansatz Materialien und Referenzwerke für den Englischunterricht für Nicht-Muttersprachler erstellt (Sinclair, 1987).

[138] Vgl. Aston (2000), auch `http://www.sslmit.unibo.it/~guy/barc.htm`.

[139] Stubbs (1996).

[140] Ein Beispiel für eine Konkordanz finden Sie in Abschnitt 5.1.2.

[141] `http://visl.sdu.dk/visl/de`.

Bühne", wie Aston es beschrieben hat. Auf den Projektseiten im Internet kann man online verschiedene Grammatikübungen in mehr als 25 Sprachen ausführen[142]. Die Übungen basieren zum Teil auf manuell vorannotierten Sätzen, zum Teil auf großen, automatisch geparsten Korpora. Dem Lerner kann dadurch eine enorme Vielfalt an authentischem Übungsmaterial angeboten werden. Neben den Syntaxübungen enthält die Seite auch eine Reihe von Sprachspielen, die sehr ansprechend aufgebaut sind. Es gibt z.B. ein kleines Fellknäuel, den *Grammar Man*, den man durch ein Labyrinth von Wortarten leiten muss, ohne einem Gespenst zu begegnen. Der richtige Weg wird jeweils durch einen Beispielsatz vorgegeben, den man aber zuerst analysieren muss. Im Hintergrund des Systems läuft ein kategorialgrammatischer Parser[143], der den Sätzen eine Dependenzanalyse zuweist.

Ein Beispiel für den zweiten Typ von Korpuseinsatz in der Fremdsprachenerwerbsforschung ist das Berliner Falko-Korpus. Im vierten Kapitel sind wir kurz auf die Annotation des Lernerkorpus eingegangen. Im Umfeld von Falko entstanden mehrere Arbeiten zum Fremdsprachenerwerb und der Didaktik von Deutsch als Fremdsprache; Maik Walter[144] zum Beispiel untersucht Satzkonnektoren wie *da, weil* oder *obwohl*, deren Verwendung gemeinhin als Indikator für die Niveaueinstufung von Lernern genutzt wird. Die Frage, ob Konnektoren tatsächlich gute Indikatoren sind, versucht Walter korpusbasiert und im Vergleich mit Daten von Muttersprachlern zu klären. Die Korpusauswertung zeigt systematische Abweichungen in der Wortstellung und der Konnektorenwahl.

Hirschmann (2015) hinterfragt die Unterscheidung von Modifikatoren und Ergänzungen sowie die Klassifizierung von Modifikatoren. Neben anderen Textsorten untersucht er den Gebrauch von Modifikatoren bei Deutschlernern im Falko-Korpus. Modalpartikeln wie *wohl, halt* oder *doch* in Sätzen wie (23) und Modalwörtern (auch: Satzadverbien, wie *wahrscheinlich, hoffentlich* oder *erfreulicherweise*) erweisen sich dabei als besonders problematisch.

(23) Sie ist wohl/halt/doch krank.

Fremdsprachenlerner verwenden diese Wortarten signifikant seltener als Muttersprachler. Modalpartikel werden teilweise in Kontexten verwendet, in denen sie nicht angemessen sind. Hirschmann entwirft ein Gesamtsystem der modifizierenden Wortarten und gliedert dabei die sonst oft als idiosynkratisch geltenden Modalpartikeln zusammen mit den Modalwörtern in das System ein.

Wir stellen im Folgenden zwei Arbeiten vor, in denen das Potenzial von Korpusanalyse und didaktischer Aufbereitung von Belegen für den Fremdsprachunterricht demonstriert wird. Es handelt sich also in beiden Fällen um Korpusarbeit *Hinter den Kulissen*.

Die erste Arbeit bezieht sich auf Modalpartikeln, die zweite auf Präpositionen. Für den Sprachlerner stellen Modalpartikeln eine besondere Herausforderung dar. Sie sind weder allein dem Lexikon noch der Grammatik zuzurechnen, ihre Funktion kann deshalb nicht einfach durch Verwendung der entsprechenden Referenzwerke erschlossen werden. Zweitens ist das komplexe Wechselspiel zwischen Partikelfunktion, Kotext und Kontext nicht leicht zu verstehen. Gerade dieses Wechselspiel kann nur anhand von au-

[142] Vgl. Bick (2005).
[143] Vgl. Karlsson (1990).
[144] Vgl. Walter (in Vorbereitung).

thentischen Beispielen vermittelt und verstanden werden[145]. Moellering begegnet diesen Problemen mit einem fremdsprachendidaktischen Programm, das auf die Verwendung authentischer Beispiele setzt. Als Materialgrundlage dienen ihr vor allem Korpora gesprochener Sprache[146], da Modalpartikeln vor allem im gesprochenen Deutsch verwendet werden. Sie ermittelt die Vorkommenshäufigkeit aller Modalpartikeln in diesen Korpora und widmet die weiteren Ausführungen den häufigsten Partikeln: *eben, nur, denn, schon, doch, mal, aber, auch, ja*. Für jede Partikel erarbeitet sie Arbeitsblätter auf der Basis von authentischen Belegen. Diese Arbeitsblätter sollen den Lernern helfen: a) die Verwendung der einzelnen lexikalischen Einheiten als Modalpartikeln von den anderen Verwendungen dieser Einheiten, z.B. als Konjunktion oder als Gradpartikel, zu unterscheiden und b) Funktion und Bedeutung der Modalpartikeln zu verstehen[147]. In Kapitel 5 dieser Arbeit werden die partikelbezogenen Lehrmaterialien vorgestellt und diskutiert. Die Materialien wurden in der Praxis erprobt, die Einstellung der Schüler zum Lernen an authentischem Sprachmaterial wurde evaluiert. Moellering sieht sich mit ihrer Arbeit in einem Trend der Fremdsprachvermittlung, die im Lehrer eher einen Vermittler als einen Wissensproduzenten sieht und Fremdsprachlernen als aktive Auseinandersetzung der Lernenden mit authentischen Äußerungen der Zielsprache[148].

Randall Jones[149] ist an Präpositionen aus der Perspektive der Fremdsprachvermittlung interessiert. Ziel seiner Studie ist es, die Beschreibungen und Lernhilfen in Lehrbüchern und Lernergrammatiken, die Präpositionen betreffen, mit den Ergebnissen der Analyse eines Korpus gesprochener Sprache zu vergleichen. Für seine Untersuchungen verwendet er ein an der Brigham Young University erstelltes Korpus des gesprochenen Deutsch (S. 118). Er betrachtet die neun am häufigsten im Korpus vorkommenden Präpositionen: *hinter, neben, zwischen, unter, vor, über, an, auf* und *in* (Tabelle 1 auf S. 120.) und stellt fest, dass eine solche Korpusanalyse andere Informationen zu Tage fördert, als sie in Sprachlehrwerken vermittelt werden. Im Detail:

- Die prototypische Unterscheidung von Ort und Richtung hilft bei der Bestimmung des Kasus, den die Präposition regiert, wenig, weil bei fast allen Präpositionen die wenigsten Vorkommen sich diesem Schema zuordnen ließen. Die meisten Vorkommen hatten keine klare lokale oder direktionale Bedeutung. Viele Präpositionen sind Teil von Präpositionalergänzungen von Verben oder Teil von idiomatischen Wendungen. In diesen Fällen ist der regierte Kasus aber nicht regelhaft erschließbar;
- die Verwendung des Akkusativs und die Verwendung des Dativs sind bei keiner Präposition ausgewogen. Bei *hinter* dominierte der Dativ mit über 80 Prozent, bei *über* der Akkusativ mit über 99 Prozent. Diese quantitativen Tendenzen zu kennen, kann für Lerner wichtig sein;

---

[145] Vgl. hierzu Moellering (2004), Kapitel 1.

[146] Zu den verwendeten Korpora s. S. 101–104. Auf S. 249 diskutiert Moellering einige Schwächen des von ihr verwendeten Korpus. Es sei erstens relativ klein und zweitens sei die überaus hohe Frequenz von *ja* dessen häufiger Verwendung als Gesprächspartikel in Telefondialogen geschuldet.

[147] Der Autorin geht es ausdrücklich nicht darum, die aktive Verwendung der Partikeln einzuüben, sondern nur darum, das Verstehen zu erleichtern, vgl. S. 244.

[148] Moellering (2004), S. 250.

[149] Vgl. Jones (2000).

- die Präpositionen selbst kommen unterschiedlich oft vor – am seltensten *hinter* und am häufigsten *in* (Tabelle 2, S. 141). Diese Erkenntnis mag vor allem für Muttersprachler banal sein, sie wird aber für den Lerner durch die Gleichbehandlung der Präpositionen in vielen Lehrbüchern verdeckt. Jones schlägt hier ein Vorgehen vom Häufigeren zum Seltneren vor.

Als Fazit schlägt Jones den verstärkten Einbezug von Korpora gesprochener und geschriebener Sprache für die Fremdsprachvermittlung oder doch zumindest für die Erstellung von Lehrwerken vor, da sie das Verständnis der komplexen Maschinerie des Deutschen erleichtern (S. 142).

Diese Arbeiten leisten einen wertvollen Beitrag zu Forschungen, die den Lernprozess nicht aus der Sicht der kognitiven Leistungen der Lernenden, sondern aus der Sicht der Besonderheiten des authentischen Sprachgebrauchs betrachten. Es bleibt zu hoffen, dass diese Erkenntnisse bei den Verlagen, die Lehrmaterialien für Deutsch als Fremdsprache erstellen, auch ankommen.

8.8 Fazit

Wie wir eingangs erwähnt haben, können die in diesem Kapitel dargestellten Arbeiten zum Teil als gute, zum Teil als schlechte Beispiele korpuslinguistischer Forschung aufgefasst werden. Wir wollen die methodischen Tendenzen, die in diesen Arbeiten deutlich werden, hier zusammenfassen und daraus Empfehlungen für ein gutes methodisches Arbeiten ableiten.

Zunächst fällt auf, dass viele Arbeiten sich auf kleinere Korpora stützen, die sich überwiegend im Besitz der Autoren befinden bzw. für diese zum Zweck der Untersuchung erstellt wurden. Es ist auch oft nicht klar, ob die Korpora digital vorliegen und maschinell ausgewertet wurden. Es ist im Prinzip nichts gegen die manuelle Auswertung eines (kleinen) Gesamtkorpus einzuwenden. Diese Methode erschwert aber die Überprüfung oder Reproduktion der Ergebnisse. Das einzige Bewertungskriterium ist in diesem Fall die Plausibilität der Ergebnisse. Die linguistische Forschung wird auch weiterhin auf Spezialkorpora angewiesen sein, die ad hoc zum Zwecke einer bestimmten Untersuchung zusammengestellt werden. Es sollte aber gefordert werden, dass diese Spezialkorpora a) digital erfasst und b) begleitend zur Publikation der Öffentlichkeit zur Verfügung gestellt werden, soweit keine urheberrechtlichen oder personenrechtlichen Gründe dagegen sprechen. Die Publikation der Daten kann entweder über die Homepage des Forschers oder über eine zentrale Sammel- und Dokumentationsstelle für Korpora geschehen. Eine solche Stelle existiert allerdings noch nicht. Auch die Beschreibung dieser Korpora mit Metadaten ist wünschenswert. Es ist erfreulich und der Konsolidierung der Korpuslinguistik als seriöse Wissenschaft förderlich, dass immer mehr Institutionen der Forschungsförderung (z.B. die Deutsche Forschungsgemeinschaft) es inzwischen als ein Förderkriterium für Forschungsarbeiten, die sich auf Korpusdaten stützen, ansieht, dass die Daten für die Öffentlichkeit verfügbar sind bzw. verfügbar gemacht werden. Andererseits entstehen Forschungsinfrastrukturen wie CLARIN (`www.clarin.eu`), die es Forschern ermöglichen, ihre projektspezifischen Daten dauerhaft zu sichern und der wissenschaftlichen Gemeinschaft zur Verfügung zu stellen.

Ein ähnliches Problem ergibt sich, wenn nicht zu wenig, sondern zu viel Daten zur Verfügung stehen. Dies ist bei Forschungen zur computervermittelten Kommunikation der Fall. Hier besteht die Tendenz, Daten in wenig kontrollierter und opportunistischer Weise zu sammeln. Auch dies erschwert letztendlich die Generalisierbarkeit der gewonnenen Erkenntnisse. Dem könnte durch den Aufbau textsorten- oder medienspezifischer Referenzkorpora abgeholfen werden. Dies ist freilich nicht die Aufgabe einzelner Wissenschaftler, sondern muss institutionell geregelt werden. Einzelne Forscher können und sollten zu einem solchen Referenzkorpus beitragen.

Es gibt nach wie vor nicht *das* Referenzkorpus des Deutschen, wie es etwa das *British National Corpus* für das Britische Englisch war und ist. Die meisten Forscher verwenden die Korpora des Instituts für Deutsche Sprache und des Digitalen Wörterbuchs der Deutschen Sprache (DWDS) an der Berlin-Brandenburgischen Akademie der Wissenschaften. Dies bedeutet auf der anderen Seite eine gewisse Verantwortung für diese Institutionen, diese Korpora permanent zur Verfügung zu stellen, zu pflegen und aktuell zu halten. Wir hoffen, mit der Darstellung der Korpuslandschaft des Deutschen in Kapitel 7 dazu beitragen zu können, dass die erwähnten Korpora stärker genutzt werden.

Es gibt kaum einen (korpus-)linguistischen Bereich oder Fragenkomplex, dem sich mehrere Arbeiten widmen. Am ehesten ist dies bisher im Bereich der Modalpartikeln geschehen. Gerade die in den vorhergehenden Kapiteln beschriebenen Probleme mit Korpusdaten als Grundlage linguistischer Erkenntnis sollten zur Reproduktion bzw. Kontrolle einmal erzielter Ergebnisse ermuntern. Verstehen Sie als Leser dieses Buches dies auch als Aufforderung, die hier beschriebenen Arbeiten und daraus gewonnenen Erkenntnisse selbst zu überprüfen.

Nicht in allen Arbeiten wird das Verhältnis von quantitativer und qualitativer Analyse reflektiert. Ein Musterbeispiel ist hier die Arbeit von Nederstigt (2003), die für alle analysierten Wörter eine die kompletten Korpusdaten umfassende quantitative Analyse vornimmt, für die darauf folgende qualitative Analyse aber für jedes Wort eine gleich große Anzahl von Belegen auswählt. Letzteres erlaubt ihr, die Analyse der beschriebenen Wörter vergleichbar zu machen. Auch die Arbeit von Peter Eisenberg zu den Anglizismen im Deutschen (Eisenberg, 2013) ist in dieser Hinsicht vorbildlich und deshalb auch unter diesem Aspekt zur Lektüre empfohlen. Die saubere Trennung beider Aspekte sollte bereits Gegenstand des Forschungsdesigns sein und vor der Auswahl der Korpora und weiteren Analysemitteln stehen.

Letztendlich müssen auch die grundsätzlichen Fragen beantwortet werden, die wir in den vorhergehenden Kapiteln aufgeworfen haben: Ist ein Korpus überhaupt geeignet zur Beantwortung der Forschungsfrage? Gibt es Alternativen oder Ergänzungen? In welchem Verhältnis stehen die ausgewählten Korpusdaten zum beschriebenen Gegenstand, sind Generalisierungen über die Korpusdaten hinaus möglich? Diese grundsätzlichen Fragen werden in den hier beschriebenen Arbeiten keinesfalls ausgeblendet, sie könnten u.E. aber stärker reflektiert werden.

8.9 Weiterführende Literatur

Es gibt mittlerweile mit dem Handbuch *Corpus Linguistics*, herausgegeben von Anke Lüdeling und Merja Kytö, eine Publikation, welche die in diesem Kapitel dargestellten korpuslinguistischen Ansätze und Themengebiete in geschlossener Form präsentiert, dies vor allem im 2009 erschienenen zweiten Band. Im Jahr 2006 ist eine *Einführung in die Korpuslinguistik* für Germanisten von Carmen Scherer erschienen. Wir denken, dass sich die Lektüre des Buches ergänzend zu diesem Buch lohnen wird. Für das Englische ist das ‚Resource Book' zur Korpuslinguistik von Tony McEnery, Richard Xiao und Yukio Tono zu empfehlen, nicht nur aber besonders auch wegen der vielen detaillierten Fallstudien in Teil C (McEnery et al., 2006).

Ansonsten ist ein regelmäßiger Blick in die Fachzeitschriften zu empfehlen. Ergiebige Quellen sind die Zeitschriften *Deutsche Sprache*, *Zeitschrift für germanistische Linguistik*, *Muttersprache*. Sie sollten außerdem die Beiträge der englischsprachigen Zeitschrift *Corpus Linguistics* und des computerlinguistisch orientierten *Journal for Language Technology and Computational Linguistics* (`http://www.jlcl.org`, Open Access) zur Kenntnis nehmen, wenn Sie up-to-date bleiben möchten. Aus der letztgenannten Zeitschrift stammt auch eine Reihe von Erfahrungsberichten über den Einsatz von Korpora und korpuslinguistischen Methoden im Universitätsunterricht, die wir vor allem den Lehrenden unter unseren Lesern ans Herz legen wollen (Beißwenger und Storrer, 2011; Bubenhofer, 2011; Dipper, 2011; Zinsmeister, 2011).

8.10 Aufgaben

1. Sie wollen untersuchen, wie oft verschiedene orthographische Varianten eines Wortes verwendet werden, oder, anders formuliert, welche Variante eines Wortes überwiegt. Sie wählen das Web als Korpus und wollen eine Suchmaschine verwenden, um anhand der gelieferten Treffer zu jeder Variante eine ungefähre Abschätzung der Verwendungshäufigkeit vorzunehmen. Arbeiten Sie mit den folgenden Beispielen: a) *Buddyliste / Buddy-Liste / Buddy Liste*, b) *Musikdownload*, *Musik-Download*, *Musik Download*, oder wählen Sie ein eigenes Beispiel. Testen Sie die Suchmaschinen Google (`www.google.de`) und Yahoo (`search.yahoo.com`). Welche Ergebnisse bringt die jeweilige Trefferliste? Prüfen Sie einige Treffer, auch solche, die weiter hinten in der Liste stehen. Sind die Treffer korrekt? Sind Sie, nach Durchsicht der Ergebnisse, der Meinung, dass eine oder mehrere der Suchmaschinen sich für solche linguistischen Untersuchungen eignen?
2. Für eine Untersuchung zu Anglizismen im Deutschen möchten Sie aus einem Korpus möglichst viele Anglizismen extrahieren. Welche Möglichkeiten sehen Sie, Anglizismen von nativen deutschen Wörtern zu unterscheiden, ohne jedes einzelne Wort zu überprüfen?
3. Bearbeiten Sie die Vorsilbe *zwischen* als Vorsilbe zu Verben wie z.B. *zwischenfinanzieren*. Suchen Sie Belege aus einem Korpus oder aus dem Web. Verfassen Sie einen Wörterbuchartikel für dieses Präfix. Erarbeiten Sie eine Übung für den Fremdsprachunterricht.

4. Betrachten Sie die E-Mail in Ihrem Postfach als eine Art Korpus. Diskutieren Sie, wenn möglich in einer Gruppe, nach welchen Textsorten Sie diese Mail sortieren könnten. Untersuchen Sie auch die Header Ihrer Mail. Welche Informationen aus dem Header lassen sich für eine Klassifikation der Nachrichten in Textsorten nutzen?
5. Auf unserer begleitenden Webseite haben wir Listen von möglichen Kollokanten für einige Schlüsselwörter bereitgestellt. Die Liste der Kollokanten wurde mit statistischen Mitteln aus einem sehr großen Korpus extrahiert. Wählen Sie aus diesen Listen alle Wortpaare aus Schlüsselwort und Kollokant aus, die Sie für die Aufnahme in ein Wörterbuch für würdig halten. Markieren Sie das Stichwort, unter dem Sie die Kollokation einordnen würden. Vergleichen Sie die Ergebnisse mit Ihren Kollegen und ermitteln Sie, wie hoch die Übereinstimmung ist. Vergleichen Sie Ihre Ergebnisse auch mit den Kollokationen in einem ein- oder zweisprachigen Wörterbuch. Welche Kollokationen zum Stichwort fallen Ihnen ein und welche Kollokationen finden Sie im Wörterbuch, die in der Liste nicht enthalten sind?
6. Eine eigene, etwas systematischere, aber keinesfalls erschöpfende Untersuchung des *Wortwarte-Korpus* förderte die folgenden Wörter mit BinnenGroßSchreibung zutage: *eBay*, *eBook*, *eGovernment*, *eLearning*, *GamerInnen*, *geWAPnet*, *LinuxTag*, *MUDder*, *WinNT*.
   Klassifizieren Sie diese Einheiten nach den Motiven, die zu diesen Bildungen führten. Fallen Ihnen weitere Beispiele ein? Nehmen Sie Stellung zu der Frage, ob die Rechtschreibnorm solche Formen zulassen sollte.

9 Glossar

Abfragesprache Eine A. ermöglicht das Suchen und Finden von Informationen in Korpora. Die gesuchten Objekte können einfache Wörter sein oder komplexe syntaktische Konstruktionen. Eine bekannte Abfragesprache ist *CQP*, eine weitere *COSMAS*, die für die Abfrage der Korpora am Institut für deutschen Sprache in Mannheim entwickelt wurde.

Alignierung In Parallelkorpora werden die Texteinheiten der Übersetzung den entsprechenden Texteinheiten des Quelltexts zugeordnet. Je nach Textsorte und Freiheit der Übersetzung, kann die A. z.B. auf Paragraphenebene stattfinden, auf Satzebene (**Satzalignierung**), auf Wortebene (**Wortalignierung**) oder z.B. bei Gedichten auch auf Versebene.

Annotation Unter A. versteht man die linguistische Anreicherung der Primärdaten eines Korpus.

Annotationsschema Ein A. ist die systematische Beschreibung von Annotationskategorien und ihre Anwendung auf Korpusdaten. Es dient als Richtlinie (**Annotationsguidelines**) beim Erstellen von annotierten Korpora und nachträglich als Dokumentation für die Annotation der erstellten Ressourcen.

Belegsammlung Eine B. ist eine Sammlung von Ausschnitten aus einem Korpus, die als Belege für ein bestimmtes linguistisches Phänomen Gegenstand weiterer linguistischer Untersuchung sind.

Generative Grammatik Als g. G. wird ein Grammatikmodell bezeichnet, nach dem durch ein begrenztes Inventar von Regeln alle wohlgeformten Sätze einer Sprache generiert werden können. Der Terminus bezeichnet außerdem die sprachwissenschaftliche Schule, in der dieses Grammatikmodell eine zentrale Rolle spielt.

Index Ein Index ist eine Liste von Wortformen, die in einem Korpus vorkommen. Die Wortformen werden zu Types zusammengefasst. Meist werden zusätzliche Informationen wie z.B. die absolute oder relative Häufigkeit des Vorkommens oder das Lemma angegeben.

Kollokation Als K. wird das wiederholte gemeinsame Vorkommen zweier Wörter in einer strukturell interessanten Einheit bezeichnet. In einer Kollokation beeinflusst ein Wort die Auswahl eines anderen Wortes zuungunsten von Wörtern mit gleicher oder ähnlicher Bedeutung

Konkordanz Eine K. ist eine Sammlung von Kotexten eines bestimmten Schlüsselworts. Kotexte einer bestimmten Länge (von Buchstaben, Wörtern oder Sätzen) um ein Schlüsselwort herum werden aus einem Korpus extrahiert und meist mit dem Schlüsselwort im Zentrum angeordnet. Konkordanzen werden vor allem bei wortbezogenen Untersuchungen verwendet.

Kontextualismus Als K. wird eine Richtung der Sprachwissenschaft bezeichnet, in der linguistische Einheiten immer im Kotext einer Äußerung und Äußerungen bzw. Texte immer im Kontext ihrer Produktion und Rezeption untersucht werden.

Kontrastives Korpus Ein k. K. enthält Texte von zwei oder mehreren Sprachen, die keine Übersetzungen voneinander sind, jedoch aus vergleichbaren Fachdomänen oder Sprachvarietäten stammen. K. K. werden vor allem für sprachvergleichende linguistische oder stilistische Untersuchungen verwendet.

Kookkurrenz Als K. wird das gemeinsame Vorkommen zweier oder mehrerer Wörter in einem Kontext von fest definierter Größe bezeichnet. Das gemeinsame Vorkommen sollte höher sein, als bei einer Zufallsverteilung aller Wörter erwartbar wäre.

Lemma Das L. ist die Grundform einer bestimmten lexikalischen Einheit und steht stellvertretend für alle Wortformen dieser lexikalischen Einheit.

Lernerkorpus In einem L. werden Äußerungen von Lernern einer Sprache gesammelt. Zusätzlich werden in den meisten Fällen zielsprachliche Normalisierungen ('Zielhypothesen') und/oder typische Lernerfehler annotiert. L. werden in der Spracherwerbsforschung und für die Sprachlehre verwendet. Ein L. der deutschen Sprache ist z.B. das Falko-Korpus an der Humboldt-Universität zu Berlin. Typischerweise dokumentieren L. den Fremdspracherwerb im Gegensatz zu Korpora des Erstspracherwerbs.

Metadaten Als M. werden Beschreibungen der Primärdaten eines Korpus bezeichnet. M. geben z.B. Auskunft über die Herkunft und den Umfang der Primärdaten.

Monitorkorpus Ein M. wird in relativ kurzen Abständen um neue Texte ergänzt, dafür werden ältere Texte entfernt. Ein M. eignet sich gut für Untersuchungen, die in kurzen Zeitabständen wiederholt werden, z.B. in der Lexikographie (Aufnahme und Beschreibung neuer Wörter und Wendungen).

Neologismus Als N. wird eine lexikalische Einheit bezeichnet, die zum Zeitpunkt der Beschreibung von vielen Sprechern als neu empfunden wird und deren Verwendung sich so weit verbreitet, dass sie in die gängigen Wörterbücher der Sprache aufgenommen wird.

Normalisierung Die N. bezeichnet allgemein eine Vereinheitlichung von Texten. Sie betrifft unterschiedliche Ebenen: (i) Zeichen- oder Dokumentkodierung z.B. Abbildung auf UNICODE; (ii) sprachlicher Ausdruck und Form z.B. in Korpora historischer Sprachstufen, von Transkripten gesprochener Sprache oder internetbasierter Kommunikation, ebenso in Lernerkorpora. Normalisierter Text kann als zusätzliche Annotationsebene vorgehalten werden oder dient als Vorverarbeitungsschritt für weiterführende Analysetools wie das Wortartentagging.

Operationalisierung Durch eine angemessene O. werden linguistische Phänomene auf Einheiten oder Relationen abgebildet, die in einem Korpus beobachtbar und wiederauffindbar sind. Sie ist die Grundlage für quantitative Auswertungen von linguistischen Konzepten und die Überprüfung von Hypothesen anhand von Korpora.

Opportunistisches Korpus Ein o. K. ist ein Korpus, welches ohne vorher festgelegte Designprinzipien danach zusammengestellt wird, welche Texte gerade verfügbar sind. O. K. sind vor allem dort angemessen, wo es allein um die Menge der Daten geht, also vor allem bei quantitativen Untersuchungen.

Paralleles Korpus Ein p. K. ist ein Korpus aus zwei oder mehr Sprachen. Die Korpustexte sind Übersetzungen von einander bzw. von einer gemeinsamen Quelle. P. K.

werden meist auf Absatz- oder Satzebene aligniert – die passenden (Ab-)Sätze werden einander zugeordnet. Parallele Korpora werden vor allem für kontrastive linguistische Studien verwendet.

Parsing Das P. bezeichnet allgemein den Prozess der syntaktischen Textanalyse. In der Psycholinguistik untersucht man das menschliche P., in der Computerlinguistik das maschinelle. Ein **Parser** ist ein Computerprogramm, das Texten eine syntaktische Analyse zuweist, z.B. in der Form eines Phrasenstruktur- oder Dependenzbaums.

Primärdaten Als P. werden die Texte bzw. Äußerungen bezeichnet, die in einem Korpus versammelt sind.

Referenzkorpus Ein R. wird als Grundlage vieler linguistischer Untersuchungen verwendet. Die Ergebnisse von Untersuchungen, die auf einem R. basieren, können so besser nachvollzogen und verglichen werden. Ein R. sollte hinsichtlich des abgebildeten Gegenstandes einen hohen Grad der Abdeckung und strukturellen Ähnlichkeit aufweisen. Kandidaten für ein Referenzkorpus der deutschen Gegenwartssprache sind die Korpora am Institut für deutsche Sprache in Mannheim und an der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin.

Tagging Beim T. werden den Token eines Korpus Wortartenlabel (so genannte **Tags**) zugeordnet. Ein Computerprogramm, das das automatisch macht, heißt **Tagger**.

Tagset Die Liste aller (morphosyntaktischen, grammatischen oder funktionalen Label, die bei einer Annotation verwendet werden.

Tokenisierung Bei der T. werden Texte in Sätze, und diese in Worttoken zerlegt. Ein **Tokenizer** ist ein Computerprogramm, das diese Zerlegung durchführt.

Vergleichskorpus Ein V. wird zur Überprüfung von Erkenntnissen verwendet, die auf Grund eines anderen Korpus gewonnen wurden. Durch das Hinzuziehen eines V. können Artefakte aufgedeckt und korrigiert werden, deren Ursache in dem für die Untersuchung verwendeten Korpus liegt.

Worttoken, Token Ein W. bezeichnet das Vorkommen eines Wortes an einer bestimmten Stelle im Korpus.

Worttype, Type In einem W. werden die Token eines Korpus zusammengefasst, die nach einem festgelegten Kriterium ähnlich oder gleich sind, z.B. Wörter mit gleicher orthographischer Form.

Literaturverzeichnis

Abney, Steven (1991): "Parsing by Chunks". In: *Principle-Based Parsing*, herausgegeben von Berwick, Robert; Abney, Steven und Tenny, Carol, Dordrecht: Kluwer Academic Publishers.

Albert, Stefanie; Anderssen, Jan; Bader, Regine; Becker, Stephanie; Bracht, Tobias; Brants, Sabine; Brants, Thorsten; Demberg, Vera; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; Hirschmann, Hagen; Janitzek, Juliane; Kirstein, Carolin; Langner, Robert; Michelbacher, Lukas; Plaehn, Oliver; Preis, Cordula; Pußel, Marcus; Rower, Marco; Schrader, Bettina; Schwartz, Anne; Smith, George und Uszkoreit, Hans (2003): *TIGER Annotationsschema, Manuskript*. Universität des Saarlandes, Universität Stuttgart, Universität Potsdam. `http://www.linguistics.ruhr-uni-bochum.de/~dipper/papers/tiger_annot.pdf`.

Allwood, Jens (2008): "Multimodal Corpora". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 207–225.

Altrichter, Helmut (2001): "Retrodigitalisierung in Deutschland – Versuch einer Zwischenbilanz". `http://www.bsb-muenchen.de/mdz/forum/altrichter/`.

Artstein, Ron und Poesio, Massimo (2008): "Inter-coder agreement for computational linguistics". *Computational Linguistics* 34 (4): S. 555–596. `https://aclweb.org/anthology/J/J08/J08-4004.pdf/`.

Aston, Guy (2000): "Learning English with the British National Corpus". In: *VI jornada de corpus lingüístics*, herausgegeben von Battaner, M.P. und Lopez, C. Barcelona, S. 15–40. `http://www.sslmit.unibo.it/~guy/barc.htm`.

Atkins, Sue; Clear, Jeremy und Ostler, Nick (1992): "Corpus Design Criteria". *Literary & Linguistic Computing* 7 (1): S. 1–16.

Augst, Gerhard (1992): "Die orthographische Integration von zusammengesetzten Anglizismen". *Sprachwissenschaft* 17: S. 45–61.

Augst, Gerhard u.a. (Herausgeber) (1997): *Zur Neuregelung der deutschen Orthographie*. Tübingen.

Baayen, Harald (2001): *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, Rolf Harald (2008): *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press. `http://www.sfs.uni-tuebingen.de/~hbaayen/publications/baayenCUPstats.pdf`.

Baroni, Marco und Bernardini, Silvia (Herausgeber) (2006): *WaCky! Working papers on the web as corpus*. Bologna: Gedit.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano und Zanchetta, Eros (2009): "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora". *Language Resources and Evaluation* 43 (3): S. 209–226.

Baroni, Marco und Evert, Stefan (2008): "Statistical Methods for Corpus Exploitation". In: *Corpus Linguistics. An International Handbook*, herausgegeben von Anke Lüdeling und Merja Kytö, Berlin: Mouton de Gruyter, S. 777–803.

Bartsch, Sabine (2002): "Anglizismen in Fachsprachen des Deutschen. Eine Untersuchung auf Basis des Darmstädter Corpus Deutscher Fachsprachen". *Muttersprache* 112 (4): S. 309–323.

Bartzsch, Rudolf (2004): *Wörterbuch überflüssiger Anglizismen*. Paderborn, 6. Auflage.

Baumann, Stefan und Riester, Arndt (2012): "Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects". In: *Prosody and meaning*, herausgegeben von Elordieta, Gorka und Prieto, Pilar, Band 25, S. 119–162.

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar und Storrer, Angelika (2012): "A TEI Schema for the Representation of Computer-mediated Communication". *Journal of the Text Encoding Initiative [Online]* (3). http://jtei.revues.org/476.

Beißwenger, Michael und Lemnitzer, Lothar (2013): "Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt ‚Digitales Wörterbuch der deutschen Sprache' (DWDS)". *JLCL* 28 (2): S. 1–22. www.jlcl.org/2013_Heft2/1BeiLem.pdf.

Beißwenger, Michael und Storrer, Angelika (2011): "Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen – Erfahrungen – Desiderate". *JLCL* 26 (1): S. 119–139. http://www.jlcl.org/2011_Heft1/9.pdf.

Bergh, Gunnar und Zanchetta, Eros (2008): "Web linguistics". In: *Corpus Linguistics. An International Handbook*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin: Mouton de Gruyter, Handbücher zur Sprache und Kommunikationswissenschaft. Volume 1, Kapitel 35, S. 309–327.

Berry, Michael W.; Drmac, Z. und Jessup, E. R. (1999): "Matrices, Vector Spaces, and Information Retrieval". *SIAM Review* 41: S. 335–362.

Biber, Douglas (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas und Jones, James K. (2009): "Quantitative Methods in Corpus Linguistics". In: *Corpus linguistics: An International Handbook*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin: Mouton de Gruyter, Handbücher zur Sprache und Kommunikationswissenschaft. Volume 2, Kapitel 61, S. 1286–1304.

Bick, Eckhard (2005): "Grammar for Fun: IT-based Grammar Learning with VISL". In: *CALL for the Nordic Languages*, herausgegeben von Henriksen, Peter Juel. Kopenhagen, Copenhagen Studies in Language, S. 49–64.

Bickerton, Derek (1984): "The language bioprogram hypothesis". *The Behavioral and Brain Sciences* 7: S. 173–188.

Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard und Zesch, Torsten (2013): "Scalable Construction of High-Quality Web Corpora". *JLCL* 28 (2): S. 23–59. www.jlcl.org/2013_Heft2/2Biemann.pdf.

Bierwisch, Manfred (1970): "Fehler-Linguistik". *Linguistic Inquiry* 1: S. 397–414.

Bird, Steven und Simons, Gary (2003): "Seven Dimensions of Portability for Language Documentation and Description". *Language* 79: S. 557–582.

Björkelund, Anders; Eckart, Kerstin; Riester, Arndt; Schauffler, Nadja und Schweitzer, Katrin (2014): "The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Re-

solution". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Island: European Language Resources Association (ELRA), S. 3222–3228. http://www.lrec-conf.org/proceedings/lrec2014/pdf/891_Paper.pdf.

Bloomfield, Leonard (1926): "A set of postulates for the science of language". *Language* 2: S. 153–164.

Boas, Hans C. und Sag, Ivan A. (Herausgeber) (2012): *Sign-based Construction Grammar*. CSLI Publications/Center for the Study of Language and Information.

Bögel, Thomas; Gertz, Michael; Gius, Evelyn; Jacke, Janina; Meister, Jan Christoph; Petris, Marco und Strötgen, Jannik (2015): "Gleiche Textdaten, unterschiedliche Erkenntnisziele? Zum Potential vermeintlich widersprüchlicher Zugänge zu Textanalyse". In: *Von Daten zu Erkenntnissen. Book of Abstracts - Vorträge*. Graz, S. 119–126. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege.

Bortz, Jürgen und Schuster, Christof (2010): *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien*. Berlin / Heidelberg / New York: Springer, 7. Auflage.

Bossong, Georg (1985): *Empirische Universalienforschung. Differentielle Objektmarkierung in der neuiranischen Sprachen*. Tübingen: Narr.

Brants, Thorsten (2000): "Inter Annotator Agreement for a German Newspaper Corpus". In: *Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athen. http://www.lrec-conf.org/proceedings/lrec2000/pdf/333.pdf.

Brants, Thorsten und Plaehn, Oliver (2000): "Interactive Corpus Annotation". In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece. http://www.lrec-conf.org/proceedings/lrec2000/pdf/334.pdf.

Breidt, Lisa (1993): "Extraction of V-N-Collocations from Text Corpora: A Feasibility Study for German". In: *Proc. Workshop on Very Large Corpora. Academic and IndustrialPerspectives. Columbus (OH)*.

Bresnan, Joan; Cueni, Anna; Nikitina, Tatiana und Baayen, Harald (2007): "Predicting the Dative Alternation". In: *Cognitive Foundations of Interpretation*, herausgegeben von Bouma, G.; Kraemer, I. und Zwarts, J., Royal Netherlands Academy of Arts and Sciences, S. 69–94.

Brill, Eric (1995): "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics* 21 (4): S. 543–565. https://aclweb.org/anthology/J/J95/J95-4004.pdf.

Bräckner, Dominik (2012): "Google Bücher aus dem Blickwinkel des Lexikographen". *Trefwoord, tijdschrift voor lexicografie* 14.

Bubenhofer, Noah (2001): "Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge". http://www.bubenhofer.com/korpuslinguistik/kurs/.

Bubenhofer, Noah (2011): "Korpuslinguistik in der linguistischen Lehre: Erfolge und Misserfolge". *JLCL* 26 (1): S. 141–156. www.jlcl.org/2011_Heft1/10.pdf.

Camp, D. De und Hancock, I. (1974): *Pidgins and creoles: Current trends and prospects*. Washington.

Carstensen, Broder und Busse, Ulrich (1993): *Anglizismen-Wörterbuch. Der Einfluß des Englischen auf den deutschen Wortschatz nach 1945*. Berlin / New York: de Gruyter.

Carstensen, Kai Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne; Klabunde, Ralf und Langer, Hagen (Herausgeber) (2010): *Computerlinguistik und Sprachtechnologie. Ei-*

ne Einführung. Elsevier, Spektrum Akademischer Verlag. 3. überarbeitete und erweiterte Auflage.

Chafe, Wallace (1992): "The importance of corpus linguistics to understanding the nature of language". In: *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, herausgegeben von Svartvik, Jan, Berlin / New York: Mouton de Gruyter, Band 65 von *Trends in Linguistics. Studies and Monographs*, S. 79–97.

Chomsky, Noam (1957): *Syntactic Structures*. Den Haag: Mouton.

Chomsky, Noam (1969): *Aspekte der Syntax-Theorie*. Frankfurt: Suhrkamp Verlag.

Chomsky, Noam (1981): *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam (1986): *Knowledge of Language*. Convergence. New York / Westport / London: Praeger.

Christ, Oliver und Schulze, B. Maximilian (1995): "Ein flexibles und modulares Anfragesystem für Textcorpora". In: *Tagungsbericht des Arbeitstreffen Lexikon + Text*. Tübingen: Niemeyer.

Claridge, Claudia (2008): "Historical Corpora". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 242–259.

Clear, Jeremy (1992): "Corpus Sampling". In: *New Directions in English Language Corpora. Methodology, Results, Software Development*, herausgegeben von Leitner, Gerhard, Berlin / New York: Narr, S. 21–31.

Cramer, Irene und Sabine Schulte im Walde im Auftrag des Instituts für Deutsche Sprache, Mannheim (Herausgeber) (2006): *Studienbibliographie Computerlinguistik und Sprachtechnologie*. Studienbibliographien Sprachwissenschaft. Tübingen: Stauffenburg Verlag Brigitte Narr GmbH. `http://www.coli.uni-saarland.de/projects/stud-bib/`.

Davies, Mark (2011): "The Corpus of Contemporary American English (COCA) and Google / Web as Corpus". `http://view.byu.edu/coca/compare-google.asp`.

den Besten, Hans und Edmondson, Jerold A. (1983): "The Verbal Complex in Continental West Germanic". In: *On the Formal Syntax of the Westgermania*, herausgegeben von Abraham, Werner, Amsterdam / Philadelphia: John Benjamins, S. 155–216.

Dern, Christa (2003): „Unhöflichkeit ist es nicht." Sprachliche Höflichkeit in Erpresserbriefen". *Deutsche Sprache* 31 (2): S. 127–141.

Diemer, Stefan (2011): "Corpus linguistics with google?" In: *Proceedings of the ISLE 2011 Conference*. Boston/Ma.

Dipper, Stefanie (2005): "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation." In: *Proceedings der Berliner XML Tage 2005 (BXML 2005)*. Berlin, S. 39–50.

Dipper, Stefanie (2008): "Theory-driven and corpus-driven computational linguistics, and the use of corpora". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 68–96.

Dipper, Stefanie (2011): "Digitale Korpora in der Lehre - Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik". *JLCL* 26 (1): S. 81–95. `www.jlcl.org/2011_Heft1/7.pdf`.

Dipper, Stefanie; Donhauser, Karin; Klein, Thomas; Linde, Sonja; Müller, Stefan und Wegera, Klaus-Peter (2013): "HiTS: ein Tagset für historische Sprachstufen des Deutschen". *JLCL* 28 (1): S. 85–137. `www.jlcl.org/2013_Heft1/5Dipper.pdf`.

Dittmann, Jürgen und Zitzke, Christine (2000): "Zur Schreibung fremdsprachlicher Komposita im Wirtschaftsdeutsch. Sprachgebrauch und neue Regelung". *Zeitschrift für ange-*

wandte Linguistik 33: S. 45–68. http://userpages.uni-koblenz.de/~diekmann/zfal/zfalarchiv/zfal33_3.pdf.

Dodd, Bill (2000): *Working with German corpora*. Birmingham: Birmingham University Press.

Doering, Nicola (2002): "Kurzm. wird gesendet' Abkürzungen und Akronyme in der SMS-Kommunikation". *Muttersprache* 112 (2): S. 97–114.

Draxler, Christoph (Herausgeber) (2008): *Korpusbasierte Sprachverarbeitung. Eine Einführung*. narr Studienbücher. Tübingen: Gunter Narr.

Dürscheid, Christa (2000a): "Rechtschreibung in elektronischen Texten". *Muttersprache* 110 (1): S. 53–62.

Dürscheid, Christa (2000b): "Verschriftlichungstendenzen jenseits der Rechtschreibreform". *Zeitschrift für germanistische Linguistik* 28: S. 223–236.

Ehrich, Veronika (2001): "Was nicht müssen und nicht können (nicht) bedeuten können: Zum Skopus der Negation bei den Modalverben des Deutschen". *Linguistische Berichte* Sonderheft 9.

Eisenberg, Peter (2013): "Anglizismen im Deutschen". In: *Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache*, herausgegeben von für Sprache und Dichtung, Deutsche Akademie und der deutschen Akademien der Wissenschaften, Union, Berlin: De Gruyter, S. 57–119.

Elsen, Hilke (2002): "Neologismen in der Jugendsprache". *Muttersprache* 112 (2): S. 136–154.

Elsen, Hilke (2004): *Neologismen. Formen und Funktionen neuer Wörter in verschiedenen Varietäten des Deutschen*. Tübingen: Narr.

Elsen, Hilke und Dzikowicz, Edyta (2005): "Neologismen in der Zeitungssprache". *Deutsch als Fremdsprache* 42 (2): S. 80–85.

Engelberg, Stefan und Lemnitzer, Lothar (Herausgeber) (2001): *Lexikographie und Wörterbuchbenutzung*, Band 14 von *Einführungen*. Tübingen: Stauffenburg.

Engelberg, Stefan und Lemnitzer, Lothar (Herausgeber) (2009): *Lexikographie und Wörterbuchbenutzung*, Band 14 von *Einführungen*. Tübingen: Stauffenburg, 4. Auflage.

Engfer, Hans-Jürgen (1996): *Empirismus vs. Rationalismus? Kritik eines philosophiegeschichtlichen Schemas*. Paderborn: Ferdinand Schöningh Verlag.

Erk, Katrin; Kowalski, Andrea und Pinkal, Manfred (2003): "A Corpus Resource for Lexical Semantics". In: *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS)*. Tilburg.

Erems, Hans-Werner und Munske, Horst Haider (Herausgeber) (1997): *Die Rechtschreibreform. Pro und Kontra*. Berlin.

Evert, Stefan (2004): "An on-line repository of association measures". http://www.collocations.de/AM.

Evert, Stefan (2006): "How random is a corpus? The library metaphor". *Zeitschrift für Anglistik und Amerikanistik* 54 (2): S. 177–190.
http://www.zaa.uni-tuebingen.de/wp-content/uploads/2006-02-Evert.pdf.

Evert, Stefan und Lüdeling, Anke (2001): "Measuring morphological productivity: Is automatic preprocessing sufficient?" In: *Proceedings of the Corpus Linguistics 2001 conference*, herausgegeben von Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew und Khoja, Shereen, S. 167–175.

Evert, Stefan und das OCWB Development Team (2010): *The CQP Query Language Tutorial (CWB version 3.0)*. Manual. http://cwb.sourceforge.net/files/CQP_Tutorial/

Fabricius-Hansen, Cathrine; Gullman, Peter; Eisenberg, Peter; Fieler, Reinhard und Peters, Jörg (Herausgeber) (2009): *Duden 4. Die Grammatik*. Mannheim: Verlag Bibliographisches Institut, 8. Auflage.

Fanselow, Gisbert (1987): *Konfigurationalität*. Tübingen: Narr.

Featherston, Sam (2007): "Data in generative grammar: the stick and the carrot". *Theoretical Linguistics* 33: S. 269–318.

Featherston, Sam (2009): "Relax, lean back, and be a linguist". *Zeitschrift für Sprachwissenschaft* 28 (1): S. 127–132.

Feine, Angelika (2003): "Fußballitis, Handyritis, Chamäleonistis. ‚-itis'-Kombinationen in der deutschen Gegenwartssprache". *Sprachwissenschaft* 28: S. 437–466.

Fellbaum, Christiane (2002): "VP idioms in the Lexicon: Topics for Research Using a Very Large Corpus". In: *Konvens 2002 – 6. Konferenz zur Verarbeitung natürlicher Sprache*. DFKI, Saarbrücken.

Fellbaum, Christiane; Kramer, Undine und Stantcheva, Diana (2004): "*Eins, einen* und *etwas* in deutschen VP-Idiomen". In: *Wortverbindungen – mehr oder weniger fest*, herausgegeben von Steyer, Kathrin, Berlin / New York: De Gruyter, S. 167–193. http://konvens2002.dfki.de/cd/pdf/fellbaum.pdf.

Fillmore, Charles J. (1968): "The Case for Case". In: *Universals in Linguistic Theory*, herausgegeben von Bach, Emmon und Harms, Robert T., Holt, Rinehart and Winston, Inc.

Fillmore, Charles (1992): "‚Corpus linguistics' or ‚computer-aided armchair linguistics' ". In: *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, herausgegeben von Svartvik, Jan, Berlin / New York: Mouton de Gruyter, Band 65 von *Trends in Linguistics. Studies and Monographs*, S. 35–60.

Firth, John Rupert (1968a): "Descriptive Linguistics and the Study of English". In: *Selected papers of J.R. Firth 1952-1959*, herausgegeben von Palmer, F.R., London: Longmans, S. 96–113.

Firth, John Rupert (1968b): "A synopsis of Linguistic Theory". In: *Selected papers of J.R. Firth 1952-1959*, herausgegeben von Palmer, F.R., London: Longmans, S. 168–205.

Firth, John Rupert (1991): "Personality and Language in Society". In: *Papers in Linguistics 1934-1951*, herausgegeben von Firth, John Rupert, London, S. 177–189.

Fischer, Rudolf-Josef (2005): *Genuszuordnung. Theorie und Praxis am Beispiel des Deutschen*. Frankfurt: Peter Lang.

Fitschen, Arne (2004): *Ein computerlinguistisches Lexikon als komplexes System*, Dissertation, Universität Stuttgart, Stuttgart. Veröffentlicht als *AIMS*, Vol 10, No. 3.

Foth, Kilian A. (2006): *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Manual. Hamburg: Fachbereich Informatik. http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204/.

Foth, Kilian A.; Köhn, Arne; Beuck, Niels und Menzel, Wolfgang (2014): "Because size does matter: The Hamburg Dependency Treebank". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Island: European Language Resources Association (ELRA), S. 2326–2333. http://www.lrec-conf.org/proceedings/lrec2014/pdf/860_Paper.pdf.

Francis, Gill (1993): "A Corpus-Driven Approach to Grammar - Principles, Methods and Examples". In: *Text and Technology. In Honour of John Sinclair*, herausgegeben von Baker,

Mona; Francis, Gill und Tognini-Bonelli, Elena, Philadelphia / Amsterdam: John Benjamins, S. 137–156.

Frank, Anette (2001): "Treebank Conversion for LTAG Grammar Extraction", presented at: Third Workshop on Linguistically Interpreted Corpora (LINC'01).

Gadamer, Hans-Georg (2010): *Gesammelte Werke: Hermeneutik: Wahrheit und Methode,-1. Grundzüge einer philosophischen Hermeneutik. Bd. 1*. Tübingen: Mohr Siebeck.

Garrapa, Luigia (2011): *Vowel Elision in Florentine Italian*. Nummer 50 in Europäische Hochschulschriften. Bern u.a.: Peter Lang.

Gee, James Paul und Grosjean, François (1983): "Performance Structures: A Psycholinguistic and Linguistic Appraisal". *Cognitive Psychology* 15: S. 411–458.

Geyken, Alexander (2007): "The DWDS corpus: a reference corpus for the German language of the twentieth century". In: *Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies*, herausgegeben von Fellbaum, Christiane, continuum, S. 23–40.

Geyken, Alexander (2011): "Die dynamische Verknüpfung von Kollokationen mit Korpusbelegen und deren Repräsentation im DWDS-Wörterbuch". *OPAL – Online publizierte Arbeiten zur Linguistik* (2): S. 9–22.

Geyken, Alexander (2013): "Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv". In: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens Altägyptisches Wörterbuch an der Berlin-Brandenburgischen Akademie der Wissenschaften*, herausgegeben von Hafemann, Ingelore, Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, S. 221–234. urn:nbn:de:kobv:b4-opus-24424.

Geyken, Alexander; Haaf, Susanne; Jurish, Bryan; Schulz, Matthias; Thomas, Christian und Wiegand, Frank (2012a): "TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv". *Jahrbuch für Computerphilologie – online*. www.computerphilologie.de/jg09/geykenetal.pdf.

Geyken, Alexander; Haaf, Susanne und Wiegand, Frank (2012b): "The dta base format: A tei-subset for the compilation of interoperable corpora". In: *11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing*, herausgegeben von Jancsary, Jeremy, Wien, S. 383–391.

Geyken, Alexander und Lemnitzer, Lothar (2012): "Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries ". In: *Proceedings of EURALEX 2012*. Oslo, S. 362–366. uw.euralex.org/elx_proceedings/Euralex2012/pp362-366%20Geyken%20and%20Lemnitzer.pdf.

Ghadessy, Mohsen; Henry, Alex und Roseberry, Robert L. (2001): *Small Corpus Studies in ELT*. Studies in Corpus Linguistics. Amsterdam / Philadelphia: John Benjamins.

Grabilovich, E. und Markovitch, S. (2007): "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis". In: *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad, Indien. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

Greenbaum, Sidney (1970): *Verb-Intensifier collocations in English – an experimental approach*. Nummer 86 in Janua Linguarum, Series minor. Den Haag: Mouton.

Greenberg, Joseph (1963): "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements". In: *Universals of Language*, herausgegeben von Greenberg, Joseph, MIT Press, S. 73–113.

Greene, B. B. und Rubin, G. M. (1971): "Automatic grammatical tagging of English". Technischer Bericht, Department of Linguistics, Brown University.

Grewendorf, Günther (1995): "Syntactic Sketches. German". In: *Syntax. Ein internationales Handbuch zeitgenössischer Forschung*, herausgegeben von Jacobs, Joachim; von Stechow, Arnim; Sternefeld, Wolfgang und Vennemann, Theo, Berlin / New York: De Gruyter, S. 1288–1319.

Gries, Stefan Th. (2008): "Dispersion and adjusted frequencies in corpora". *International Journal of Corpus Linguistics* 13 (4): S. 403–437.

Gupta, Piklu (2000): "German be-verbs revisited: using corpus evidence to investigate valency". In: *Working with German corpora*, herausgegeben von Dodd, Bill, Birmingham: Birmingham University Press, S. 96–115.

Haaf, Susanne; Wiegand, Frank und Geyken, Alexander (2013): "Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text." *Journal of the TEI – online* 4. `https://jtei.revues.org/739`.

Haase, Martin; Huber, Michael; Krumeich, Alexander und Rehm, Georg (1997): "Internetkommunikation und Sprachwandel". In: *Sprachwandel durch Computer*, herausgegeben von Weingarten, Rüdiger, Opladen, S. 51–85.

Haider, Hubert (1985): "The case of German". In: *Studies in German grammar*, herausgegeben von Toman, Jindřich, Dordrecht: Foris, S. 65–101.

Halliday, M.A.K. (1992): "Language as system and language as instance: The corpus as a theoretical construct". In: *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, herausgegeben von Svartvik, Jan, Berlin / New York: Mouton de Gruyter, Band 65 von *Trends in Linguistics. Studies and Monographs*, S. 61–77.

Harris, Randy Allen (1995): *The linguistics wars*. Oxford: Oxford Univ. Press.

Harris, Zellig S. (1951): *Methods in Structural Linguistics*. Chicago: University of Chicago Press. Neuaufgelegt als *Structural Linguistics*, 1960.

Hausmann, Franz Josef (1985): "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels". In: *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.6 1984*, herausgegeben von Bergenholtz, Henning und Mugdan, Joachim, Tübingen: Niemeyer, S. 118–129.

Hausmann, Franz Josef (2004): "Was sind eigentlich Kollokationen?" In: *Wortverbindungen - mehr oder weniger fest. Jahrbuch 2003 des Instituts für deutsche Sprache*, herausgegeben von Steyer, Kathrin, Berlin / New York, S. 309–334.

Helbig, Gerhard (1994): *Lexikon deutscher Partikeln*. Leipzig: Langenscheidt.

Herberg, Dieter; Kinne, Michael und Steffens, Doris (2004): *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. Berlin: De Gruyter.

Hinrichs, Erhard; Kübler, Sandra; Naumann, Karin; Telljohann, Heike und Trushkina, Julia (2004): "Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank". In: *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT)*.

Hirschmann, Hagen (2015): *Modifikatoren im Deutschen. Ihre Klassifizierung und varietätenspezifische Verwendung*. Nummer 86 in Studien zur deutschen Grammatik. Tübingen: Stauffenburg.

Hjelmslev, Louis (1974): *Prolegomena zu einer Sprachtheorie*, Band 9 von *Linguistische Reihe*. München: Hueber.

Hockett, Charles F. (1964): "Sound Change". *Language* 41: S. 185–204.

Hohle, Tilmann N. (1986): "Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder". In: *Akten des VII. Internationalen Germanisten-Kongresses Göttingen 1985*, Tübingen: Niemeyer, Band 3, S. 329–340.

Hovy, Eduard; Marcus, Mitchell; Palmer, Martha; Ramshaw, Lance und Weischedel, Ralph (2006): "OntoNotes: The 90% Solution". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, S. 57–60. http://www.aclweb.org/anthology/N09-4006.

Hundt, Marianne; Nesselhauf, Nadja und Biewer, Carolin (2007): *Corpus Linguistics and the Web*. Amsterdam / New York: Rodopi.

Hundt, Markus (2006): "Deutschsprachige Einführung in die Korpuslinguistik. Rezension zu Lothar Lemnitzer / Heike Zinsmeister, Korpuslinguistik. Eine Einführung. Tübingen: Narr 2006". *Sprachreport* 4: S. 19–22.

Hunston, Susan (2008): "Collection strategies and design decisions". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 154–168.

Ide, Nancy und Suderman, Keith (2007): "GrAF: A graph-based format for linguistic annotations". In: *Proceedings of the Linguistic Annotation Workshop*. Prague, Czech Republic: Association for Computational Linguistics, S. 1–8. www.aclweb.org/anthology/W07-1501.

Ivanova, Kremena; Heid, Ulrich; Schulte im Walde, Sabine; Kilgarriff, Adam und Pomikálek, Jan (2008): "Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case". In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, S. 2101–2107. http://www.lrec-conf.org/proceedings/lrec2008/pdf/537_paper.pdf.

Jackendoff, Ray S. (1977): *$\overline{X}$ Syntax: A Study of Phrase Structure*. Cambridge: Massachusetts, London: England: The MIT Press.

Johnson, Keith (2008): *Quantitative Methods in Linguistics*. Oxford: Blackwell Publishing.

Jones, Randall L. (2000): "A corpus-based study of German accusative/dative prepositions". In: *Working with German corpora*, herausgegeben von Dodd, Bill, Birmingham: Birmingham University Press, S. 116–142.

Jurafsky, Daniel S. und Martin, James H. (2000): *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Jurafsky, Daniel S. und Martin, James H. (2008): *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2. Auflage.

Jurish, Bryan (2013): "Canonicalizing the Deutsches Textarchiv". In: *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens Altägyptisches Wörterbuch an der Berlin-Brandenburgischen Akademie der Wissenschaften*, herausgegeben von Hafemann, Ingelore. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, S. 235–244. urn:nbn:de:kobv:b4-opus-24433.

Jurish, Bryan; Thomas, Christian und Wiegand, Frank (2014): "Querying the Deutsches Textarchiv". In: *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities*, herausgegeben von Kruschwitz, U.; Hopfgartner, F. und Gurrin, C. Berlin, S. 25–30. http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

Jurish, Bryan und Würzner, Kay-Michael (2013): "Word and Sentence Tokenization with Hidden Markov Models". *JLCL* 28 (2): S. 61–83. www.jlcl.org/2013 Heft2/3Jurish.pdf.

Kamp, Hans und Reyle, Uwe (1993): *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Karlsson, Fred (1990): "Constraint grammar as a framework for parsing running text". In: *Papers presented to the 13th International Conference on Computational Linguistics*, herausgegeben von Karlgren, Hans. Helsinki, Band 3, S. 168–173.

Karlsson, Fred (2008): "Early generative linguistics and empirical methodology". In: *Corpus Linguistics. An International Handbook. Volume 1*. herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 14–32.

Keil, Martina (1997): *Wort für Wort: Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*. Nummer 35 in Sprache und Information. Tübingen: Niemeyer.

Kenny, Dorothy (2000): "Translators at play: exploitations of collocational norms in German-English translation". In: *Working with German corpora*, herausgegeben von Dodd, Bill, Birmingham: Birmingham University Press, S. 143–160.

Kepser, Stefan und Reis, Marga (2008): *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*. Berlin / Boston: De Gruyter Mouton.

Kermes, Hannah (2003): *Off-line (and On-line) Text Analysis for Computational Lexicography*. Dissertation, Universität Stuttgart, Stuttgart. Veröffentlicht als *AIMS*, Vol 9, No. 3.

Kertész, András und Rákosi, Csilla (2012): *Data and Evidence in Linguistics. A Plausible Argumentation Model*. Cambridge: CUP.

Kilgarriff, Adam (2007): "Googleology is Bad Science". *Computational Linguistics* 33 (1): S. 147–151. https://aclweb.org/anthology/J/J07/J07-1010.pdf.

Kilgarriff, Adam und Grefenstette, Gregory (2003): "Introduction to the special issue on the web as corpus". *Computational Linguistics* 29 (3): S. 333–347. https://aclweb.org/anthology/J/J03/J03-3001.pdf.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel und Tugwell, David (2004): "The Sketch Engine". In: *Proceedings of EURALEX 2004*. Lorient, S. 105–116. https://aclweb.org/anthology/J/J03/J03-3001.pdf.

Kiss, Tibor (2011): "Bedingungen für den Wegfall eines Artikels: Distribution und Interpretation von Präposition-Nomen-Kombinationen". In: *Sprachliches Wissen zwischen Lexikon und Grammatik (= Jahrbuch des Instituts für deutsche Sprache)*, herausgegeben von Engelberg, Stefan; Holler, Anke und Proost, Kristel, Berlin / New York: Walter de Gruyter, S. 251–283.

Kiss, Tibor; Müller, Antje; Roch, Claudia; Stadtfeld, Tobias; Börner, Katharina und Duzy, Monika (2014): "Ein Handbuch für die Bestimmung und Annotation von Präpositionsbedeutungen im Deutschen". *Bochumer Linguistische Arbeitsberichte* 14. http://www.linguistics.ruhr-uni-bochum.de/bla/014-kiss_etal2014.pdf.

Klenk, Ursula (2003): *Generative Syntax*. narr studienbücher. Tübingen: Narr.

Klosa, Annette (2003): "gegen-Verben – ein neues Wortbildungsmuster". *Sprachwissenschaft* 28: S. 467–494.

Kniffka, Gabriele (1996): *NP Aufspaltung im Deutschen*. Kölner linguistische Arbeiten – Germanistik; 31. Hürth: Gabel.

Koch, Peter und Oesterreicher, Wulf (1994): "Schriftlichkeit und Sprache". In: *Schrift und Schriftlichkeit*, herausgegeben von Günther, H. und Ludwig, O., Berlin / New York: De Gruyter, Band 1 von *Handbücher für Sprach- und Kommunikationswissenschaft*, S. 587–604.

König, Ekkehard; Stark, Detlef und Requardt, Susanne (Herausgeber) (1990): *Adverbien und Partikeln: ein deutsch-englisches Wörterbuch*. Heidelberg: Groos.

Krasselt, Julia; Bollmann, Marcel; Dipper, Stefanie und Petran, Florian (2015): "Guidelines für die Normalisierung historischer deutscher Texte / Guidelines for Normalizing Historical German Texts". *Bochumer Linguistische Arbeitsberichte* 15. `www.linguistics.ruhr-uni-bochum.de/bla/015-krasselt_etal2015.pdf`.

Kübler, Sandra und Zinsmeister, Heike (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.

Kučera, Henry und Francis, Nelson W. (1967): *Computational Analysis of Present-day American English*. Providence: Brown University Press.

Labov, William (1975): *What is a Linguistic Fact?* Lisse: The Peter de Ridder Press.

Landauer, T. K. und Dumais, S. T. (1997): "A Solution to Plato's Problem. The Latent Semantic Analysis theory of the acquisition, induction and representation of knowledge". *Psychological Review* 104 (2): S. 211–240.

Landauer, T. K.; Foltz, P. W. und Laham, D. (1998): "Introduction to Latent Semantic Analysis". *Discourse Processes* 25: S. 259–284.

Langer, Hagen (2010): "Syntax und Parsing". In: *Computerlinguistik und Sprachtechnologie. Eine Einführung*, herausgegeben von Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia; Jekat, Susanne; Klabunde, Ralf und Langer, Hagen, Spektrum Akademischer Verlag, S. 280–329. 3. Auflage.

Langner, Helmut (2001): "Zum Wortschatz der Sachgruppe Internet". *Muttersprache* 111 (2): S. 97–109.

Larson-Hall, Jenifer (2010): *A Guide to Doing Statistics in Second Language Research using SPSS*. New York / London: Routledge.

Leech, Geoffrey (1992): "Corpora and theories of linguistic performance". In: *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82*, herausgegeben von Svartvik, Jan, Berlin / New York: Mouton de Gruyter, Band 65 von *Trends in Linguistics. Studies and Monographs*, S. 105–122.

Leech, Geoffrey (1997): "Introducing Corpus Annotation". In: *Corpus Annotation. Linguistic Information from Computer Text Corpora*, herausgegeben von Garside, Roger; Leech, Geoffrey und McEnery, Tony, London / New York: Longman, S. 1–18.

Leech, Geoffrey und Wilson, Andrew (1996): "EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora". Technischer Bericht, Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-TCWG-MAC/R. `www.ilc.cnr.it/EAGLES/annotate/annotate.html`.

Lehmberg, Timm; Rehm, Georg; Witt, Andreas und Zimmermann, Felix (2008): "Digital text collections, linguistic research data, and mashups: Notes on the legal situation". *Library Trends* 57 (1): S. 52–71.

Lehr, Andrea (1996): *Kollokationen in maschinenlesbaren Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze*, Band 168 von *RGL*. Tübingen: Niemeyer.

Lemnitzer, Lothar (1997): *Extraktion komplexer Lexeme aus Textkorpora*. Tübingen: Niemeyer.

Lemnitzer, Lothar (2001): "Wann kommt er denn nun wohl endlich zur Sache? Modalpartikel-Kombinationen. Eine korpusbasierte Untersuchung". In: *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik*, herausgegeben von et al., Andreo Lehr, Berlin / New York, S. 349–371.

Lemnitzer, Lothar (2013): "Making sense of nonce words ". In: *Nye Ord*, herausgegeben von Andersen, Margrethe Heidemann und Jensen, Joergen Noerby, Kopenhagen, S. 7–18.

Lemnitzer, Lothar und Geyken, Alexander (2014): "Extraktion lexikographischer Informationen aus Textkorpora". In: *Internetlexikographie*, herausgegeben von Klosa, Annette, Berlin: De Gruyter.

Lemnitzer, Lothar und Naumann, Karin (2001): "„Auf Wiederlesen !" – das schriftlich verfaßte Unterrichtsgespräch in der computervermittelten Kommunikation. Bericht von einem virtuellen Seminar". In: *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermitteiter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*, herausgegeben von Beißwenger, Michael, Stuttgart: ibidem, S. 469–491.

Leuninger, Helen (1996): *Reden ist Schweigen, Silber ist Gold. Gesammelte Versprecher*. München: dtv.

Levin, Beth (1993): *English Verb Classes and Alternations*. Chicago: The University of Chicago Press.

Lezius, Wolfgang (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Dissertation, Universität Stuttgart, Stuttgart. Veröffentlicht als *AIMS*, Vol. 8, No. 4. `http://www.wolfganglezius.de/lib/exe/fetch.php?media=cl:disslezius.pdf`.

Lichte, Timm (2005): "Corpus-based Acquisition of Complex Negative Polarity Items". In: *Proceedings of the Tenth ESSLLI Student Session*. Edinburgh.

Lin, Yuri; Michel, Jean-Baptiste; Aiden, Erez Lieberman; Orwant, Jon; Brockman, Will und Petrov, Slav (2012): "Syntactic Annotations for the Google Books NGram Corpus". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, herausgegeben von for Computational Linguistics, Association, Jeju, S. 169–174. `https://aclweb.org/anthology/P/P12/P12-3029.pdf`.

Lobin, Henning (2000): *Informationsmodellierung in XML und SGML*. Berlin: Springer.

Lüdeling, Anke und Evert, Stefan (2003): "Linguistic experience and productivity: Corpus evidence for fine-grained distinctions". In: *Proceedings of the Corpus Linguistics 2003 conference*, herausgegeben von Archer, Dawn; Rayson, Paul; Wilson, Andrew und McEnery, Tony, S. 475–483.

Lüdeling, Anke und Evert, Stefan (2004): "The emergence of productive non-medical *-itis*: Corpus evidence and qualitative analysis". In: *Proceedings of the First International Conference on Linguistic Evidence*.

Lüdeling, Anke; Evert, Stefan und Heid, Ulrich (2000): "On Measuring Morphological Productivity". In: *KONVENS-2000 – Sprachkommunikation*, herausgegeben von Schukat-Talamazzini, Ernst G. und Zühlke, Werner, S. 215–220.

Lüdeling, Anke und Kytö, Merja (Herausgeber) (2008): *Corpus Linguistics. An International Handbook. Volume 1*. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: De Gruyter.

Lüdeling, Anke und Kytö, Merja (Herausgeber) (2009): *Corpus Linguistics. An International Handbook. Volume 2*. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: De Gruyter.

Lüdeling, Anke; Poschenrieder, Thorwald und Faulstich, Lukas C. (2005a): "DeutschDiachronDigital – ein diachrones Korpus des Deutschen". In: *Jahrbuch für Computerphilologie 2004*, herausgegeben von Georg Braungart, Peter Gendolla, Fotis Jannidis. mentis Verlag. http://www.informatik.hu-berlin.de/Forschung_Lehre/wbi/publications/2005/ddd-computerphilologie.pdf.

Lüdeling, Anke und Walter, Maik (2010): "Korpuslinguistik". In: *Handbuch Deutsch als Fremd- und Zweitsprache*, herausgegeben von Krumm, Hans-Jürgen; Fandrych, Christian; Hufeisen, Britta und Riemer, Claudia. Berlin: Mouton de Gruyter.

Lüdeling, Anke; Walter, Maik; Kroymann, Emil und Adolphs, Peter (2005b): "Multi-Level Error Annotation in Learner Corpora". In: *Proceedings of the Corpus Linguistics 2005*. Birmingham.

Maden-Weinberger, Ursula (2008): "Modality as Indicator of L2 Proficiency? A corpus-based investigation into advanced German interlanguage". In: *Fortgeschrittene Lernervarietäten und Zweitspracherwerbsforschung*, herausgegeben von Walter, Maik und Grommes, Patrick. Berlin: De Gruyter, S. 141–164.

Mahlberg, Michaela und Brook O'Donnell, Matthew (2010): *Terms in Corpus Linguistics*. London: continuum.

Mann, William C. und Thompson, Sandra A. (1988): "Rhetorical Structure Theory: Toward a functional theory of text organization". *Text* 8 (3): S. 243–281.

Manning, Christopher D. und Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing*. Cambridge (Mass.) and London: The M.I.T. Press.

Marcus, Mitchell; Kim, Grace; Marcinkiewicz, Mary Ann; MacIntyre, Robert; Bies, Ann; Ferguson, Mark; Katz, Karen und Schasberger, Britta (1994): "The Penn treebank: Annotating predicate argument structure". In: *ARPA Human Language Technology Workshop*.

Marcus, Mitchell P.; Santorini, Beatrice und Marcinkiewicz, Mary Ann (1993): "Building a large annotated corpus of English: the Penn Treebank". *Computational Linguistics* 19: S. 313–330. https://aclweb.org/anthology/J/J93/J93-2004.pdf.

McEnery, Tony und Hardie, Andrew (2012): *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony und Wilson, Andrew (1996): *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

McEnery, Tony und Wilson, Andrew (2001): *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press, 2. Auflage.

McEnery, Tony; Xiao, Richard und Tono, Yukio (2006): *Corpus-Based Language Studies. An advanced resource book*. Routledge Applied Linguistics. London: Routledge.

Meindl, Claudia (2011): *Methodik für Linguisten: eine Einführung in Statistik und Versuchsplanung*. Tübingen: Narr.

Meurers, W. Detmar (2005): "On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German". *Lingua* 115: S. 1619–1639.

Meurers, W. Detmar und Müller, Stefan (2008): "Corpora and Syntax". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja. Berlin: De Gruyter, S. 920–933.

Meyer, Markus (2009): "Sprachliche Wohlgeformtheit - eine kritische Bestandsaufnahme". *Zeitschrift für Sprachwissenschaft* 28 (1): S. 141–150.

Mindt, Dieter (1996): "English corpus linguistics and the foreign language teaching syllabus". In: *Using corpora for language research. Studies in the honour of Geoffrey Leech*, herausgegeben von Thomas, Jenny und Short, Mick, London: Longman, S. 232–248.

Mitkov, Ruslan; Evans, Richard; Orasan, Constantin; Barbu, Catalina; Jones, Lisa und Sotirova, Violeta (2000): "Coreference and Anaphora: Developing Annotating Tools, Annotated Resources and Annotation Strategies". In: *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*. Lancaster.

Moellering, Martina (2004): *The Acquisition of German Modal Particles. A Corpus-based Approach*, Band 10 von *Linguistic Insights. Studies in Language and Communication*. Bern: Peter Lang.

Mukherjee, Joybrato (2002): *Korpuslinguistik und Englischunterricht: eine Einführung*, Band 14 von *Sprache im Kontext*. Frankfurt: Peter Lang.

Mukherjee, Joybrato (2009): *Anglistische Korpuslinguistik: Eine Einführung*. Grundlagen der Anglistik und Amerikanistik 33. Berlin: Erich Schmidt.

Müller, Antje (2013): *Spatiale Bedeutungen deutscher Präpositionen. Bedeutungsdifferenzierung und Annotation*. Dissertation, Bochum:. Veröffentlicht als Bochumer Linguistische Arbeitsberichte (BLA) 11. `http://www.linguistics.ruhr-uni-bochum.de/bla/011-mueller2013.pdf`

Müller, Frank H. (2004): *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Universität Tübingen. `http://www.sfs.uni-tuebingen.de/tupp/doc/stylebook.pdf`.

Müller, Stefan (2003): "Mehrfache Vorfeldbesetzung". *Deutsche Sprache* 31 (1): S. 29–62. `http://hpsg.fu-berlin.de/~stefan/PS/vorfeld-ds2003.pdf`.

Müller, Stefan (2005): "Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung". *Linguistische Berichte* 203: S. 297–330. `https://hpsg.fu-berlin.de/~stefan/Pub/mehr-vf-lb.html`.

Naumann, Karin (2005): *Manual for the Annotation of in-document Referential Relations*. Universität Tübingen. `http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-coreference-manual-2007.pdf`.

Nederstigt, Ulrike (2003): Auch *and* noch *in child and adult German*, Band 23 von *Studies in Language Acquisition*. Berlin / New York: Mouton de Gruyter.

Nesselhauf, Nadja (2004): "Learner Corpora and their Potential for Language Teaching". In: *How to Use Corpora in Language Teaching*, herausgegeben von Sinclair, John, Amsterdam: John Benjamins, S. 125–152.

Nivre, Joakim (2008): "Treebanks". In: *Corpus Linguistics. An International Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 225–241.

Nübling, Damaris und Szczepaniak, Renata (2011): "Merkmal(s?)analyse, Seminar(s?)arbeit und Essen(s?)ausgabe: Zweifelsfälle der Verfugung als Indikatoren für Sprachwandel". *Zeitschrift für Sprachwissenschaft* 30: S. 45–73. `http://tinyurl.com/oh72ky3`.

Oakes, Michael P. (1998): *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

Ooi, Vincent B.Y. (1998): *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Ossner, Jakob und Zinsmeister, Heike (2014): *Sprachwissenschaft für das Lehramt*. Paderborn: Ferdinand Schöningh.

Palmer, Martha; Gildea, Dan und Kingsbury, Paul (2005): "The Proposition Bank: A Corpus Annotated with Semantic Roles". *Computational Linguistics* 31 (1). https://aclweb.org/anthology/J/J05/J05-1004.pdf.

Paprotté, Wolf (1992): "Korpuslinguistik - Rückkehr zum Strukturalismus oder Erneuerung der Computerlinguistik?" *LDV-Forum* 9.2: S. 3–14.

Paprotté, Wolf (1994): "Theorie und Empirie in der Linguistik: Neue Wege der Korpuslinguistik". In: *Satz – Text – Diskurs. Akten des 27. Linguistischen Kolloquiums, Münster 1992*, herausgegeben von Beckmann, Susanne und Frilling, Sabine. Tübingen: Niemeyer, Band 2, S. 19–26.

Perkuhn, Rainer; Keibel, Holger und Kupietz, Marc (2012): *Korpuslinguistik*. Paderborn: Wilhelm Fink.

Peschel, Corinna (2002): *Zum Zusammenhang von Wortneubildung und Textkonstitution*, Band 237 von *RGL*. Tübingen: Niemeyer.

Pittner, Karin (1999): *Adverbiale im Deutschen. Untersuchungen zu ihrer Stellung und Interpretation*. Studien zur deutschen Grammatik 60. Tübingen: Stauffenburg.

Pittner, Karin und Berman, Judith (2013): *Deutsche Syntax. Ein Arbeitsbuch*. Narr Studienbücher. Tübingen: Narr, 5. Auflage.

Poesio, Massimo (2004): "Coreference". MATE Dialogue Annotation Guidelines-Deliverable 2.1, S. 126–182. http://www.andreasmengel.de/pubs/ndag.pdf.

Poesio, Massimo und Vieira, Renata (1998): "A Corpus-based Investigation of Definite Description Use". *Computational Linguistics* 24 (2): S. 183–216. https://aclweb.org/anthology/J/J98/J98-2001.pdf.

Poethe, Hannelore (2000): "Wortbildung und Orthographie". *Muttersprache* 110 (1): S. 37–51.

Prince, Ellen F. (1981): "Toward a taxonomy of given-new information". In: *Radical Pragmatics*, herausgegeben von Cole, Peter, New York: Academic Press, S. 223–255.

Prince, Ellen F. (1992): "The ZPG Letter: Subjects, Definiteness, and Information-status". In: *Discourse Description: Diverse Analyses of a Fund Raising Text*, herausgegeben von Mann, William C. und Sandra A. Thompson, Amsterdam / Philadelphia: John Benjamins Publishing Company, S. 295–325.

Pullum, Geoffrey K. (1991): *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language*. Chicago: The University of Chicago Press.

Pullum, Geoffrey K. (2003): "Corpus fetishism". Language Log, 16. Nov. 2003. http://itre.cis.upenn.edu/~myl/languagelog/archives/000122.html.

Pusch, Luise (1984): "Sie sah zu ihm auf wie zu einem Gott. Das Duden-Bedeutungswörterbuch als Trivialroman". In: *Das Deutsche als Männersprache*, herausgegeben von Pusch, Luise, Frankfurt/M.: Suhrkamp, S. 135–144.

Pustejovsky, James und Stubbs, Amber (2012): *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Quasthoff, Uwe (Herausgeber) (2007): *Deutsches Neologismenwörterbuch. Neue Wörter und Wortbedeutungen in der Gegenwartssprache*. Berlin: De Gruyter.

Rehbein, Ines (2010): "Der Einfluss der Dependenzgrammatik auf die Computerlinguistik". *Zeitschrift für Germanistische Linguistik (ZGL)* 38 (2): S. 224–248.

Reznicek, Marc; Lüdeling, Anke und Hirschmann, Hagen (2013): "Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture". In: *Automatic Treat-*

ment and Analysis of Learner Corpus Data, herausgegeben von Díaz-Negrillo, Ana; Ballier, Nicolas und Thompson, Paul, Amsterdam: John Benjamins, S. 101–123.

Reznicek, Marc; Lüdeling, Anke; Krummes, Cedric; Schwantuschke, Franziska; Walter, Maik; Schmidt, Karin; Hirschmann, Hagen und Andreas, Torsten Andreas (2012): *Das Falko-Handbuch. Korpusaufbau und Annotationen*. Humboldt-Universität zu Berlin, 2. Auflage. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/.

Reznicek, Marc und Zinsmeister, Heike (2013): "STTS-Konfusionsklassen beim Tagging von Fremdsprachlernertexten". *JLCL* 28 (1): S. 63–83. http://www.jlcl.org/2013_Heft1/4Reznicek.pdf.

Riehemann, Susanne (1993): "Word Formation in Lexical Type Hierarchies – A Case Study of bar-Adjectives in German". SfS Report 2-93, Seminar für Sprachwissenschaft, Eberhard-Karls-Universität Tübingen.

Ruge, Nikolaus (2004): "Das Suffixoid ‚-technisch' in der Wortbildung der deutschen Gegenwartssprache". *Muttersprache* 114 (1): S. 29–41.

Runkehl, Jens; Schlobinski, Peter und Siever, Torsten (1998): "Sprache und Kommunikation im Internet". https://www.mediensprache.net/de/literatur/show.aspx?id=2.

Ruppenhofer, Josef; Ellsworth, Michael; Petruck, Miriam RL; Johnson, Christopher R und Scheffczyk, Jan (2006): "FrameNet II: Extended theory and practice". Technischer Bericht, International Computer Science Institute, Berkeley, CA. http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.

Sampson, Geoffrey (1996): "From central embedding to corpus linguistics". In: *Using corpora for language research. Studies in the honour of Geoffrey Leech*, herausgegeben von Thomas, Jenny und Short, Mick, London: Longman, S. 14–26.

Sampson, Geoffrey (2003): "Thoughts of Two Decades of Drawing Trees". In: *Treebanks. Building and Using Parsed Corpora*, herausgegeben von Abeillé, Anne, Kluwer Academic Publisher, S. 23–41.

Sampson, Geoffrey und McCarthy, Diana (2004): *Corpus Linguistics: Readings in a Widening Discipline*. Open Linguistics Series. Continuum.

Sarasin, Philipp (2012): "Sozialgeschichte vs. Foucault im Google Books NGram Viewer. Ein alter Streitfall in einem neuen Tool". In: *Wozu noch Sozialgeschichte? Eine Disziplin im Umbruch*, herausgegeben von Maeder, Pascal; Lüthi, Barbara und Mergel, Tomas, Göttingen: Vandenhoek und Ruprecht, S. 151–174.

Sasaki, Felix und Witt, Andreas (2004): "Linguistische Korpora". In: *Texttechnologie – Perspektiven und Anwendungen*, herausgegeben von Lobin, Henning und Lemnitzer, Lothar, Tübingen: Stauffenburg, S. 13–49. Mit einem Exkurs von Eva Anna Lenz.

Sasano, Ryohei; Kawahara, Daisuke und Kurahashi, Sadaho (2009): "The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis". In: *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL.*, Boulder/Colorado, S. 521–529. https://aclweb.org/anthology/N/N09/N09-1059.pdf.

Schade, Ulrich; Barattelli, Stefan; Lingnau, Beate; Hadelich, Kerstin und Dipper, Stefanie (2003): "Relativsatzproduktion". *Linguistische Berichte* 193: S. 33–53.

Scherer, Carmen (2005): *Wortbildungswandel und Produktivität. Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Schiller, Anne; Teufel, Simone; Stöckert, Christine und Thielen, Christine (1999): "Guidelines für das Tagging deutscher Textcorpora mit STTS". Technischer Bericht, Institut für ma-

schnelle Sprachverarbeitung, Stuttgart. http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf.

Schmid, Helmut (1995): "Improvements in part-of-speech tagging with an application to German". In: *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, Helmut (2008): "Tokenizing and Part-of-Speech Tagging". In: *Corpus Linguistics. An international Handbook. Volume 1*, herausgegeben von Lüdeling, Anke und Kytö, Merja, Berlin, S. 527–551.

Schmid, Helmut und Laws, Florian (2008): "Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging". In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK, S. 777–784. https://aclweb.org/anthology/C/C08/C08-1098.pdf.

Schmidt, Ingrid (2004): "Modellierung von Metadaten". In: *Textrechnologie – Perspektiven und Anwendungen*, herausgegeben von Lobin, Henning und Lemnitzer, Lothar, Tübingen: Stauffenburg, S. 143–164.

Schmidt, Thomas (2005): "Modellbildung und Modellierungsparadigmen in der computergestützten Korpuslinguistik". In: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*, herausgegeben von Fisseni, Bernhard; Schmitz, Hans-Christian; Schröder, Bernhard und Wagner, Petra, Frankfurt/M.: Peter Lang, S. 290–301.

Schmidt, Thomas und Wörner, Kai (2012): *Multilingual Corpora and Multilingual Corpus Analysis*, Band 14 von *Hamburg Studies in Multilingualism*. Amsterdam: John Benjamins.

Scholze-Stubenrecht, Werner (2002): "‚Die Auswahl der Einträge ist äußerst beliebig.' Warum *Jagdherr* und *Pokémon* nicht im Duden stehen". *Sprachwissenschaft* 27: S. 225–248.

Schulte im Walde, Sabine (2003): *Experiments on the Automatic Induction of German Semantic Verb Classes*. Dissertation, Universität Stuttgart. Published as AIMS Report 9(2). http://www.schulteimwalde.de/research/phd.html

Schulte im Walde, Sabine und Müller, Stefan (2013): "Using Web Corpora for the Automatic Acquisition of Lexical-Semantic Knowledge". *JLCL* 28 (2): S. 85–105. http://www.jlcl.org/2013_Heft2/4siw-mueller.pdf.

Schwitalla, Johannes (2002): "Kleine Botschaften. Telegramm- und SMS-Texte". *Osnabrücker Beiträge zur Sprachtheorie* (64): S. 33–56.

Sharoff, Serge (2006): "Creating general-purpose corpora using automated search engine queries". In: *WaCky! Working papers on the web as corpus*, herausgegeben von Baroni, Marco und Bernardini, Silvia, Bologna: Gedit. Corpora und Abfrage: http://corpus.leeds.ac.uk/internet.html.

Silverman, Kim; Beckman, Mary; Pitrelli, John; Ostendorf, Mari; Wightman, Colin; Price, Patti; Pierrehumbert, Janet und Hirschberg, Julia (1992): "TOBI: a standard for labeling English prosody." In: *The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16, 1992*. http://www.isca-speech.org/archive/icslp_1992/i92_0867.html.

Simov, Kiril und Osenova, Petya (2003): "Practical Annotation Scheme for an HPSG Treebank of Bulgarian". In: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-2003)*. Budapest, S. 17–24. https://aclweb.org/anthology/W/W03/W03-2403.pdf.

Sinclair, John (Herausgeber) (1987): *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John (1996): "EAGLES Preliminary recommendations on Corpus Typology". `http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html`.

Starke, Günter (1993): "Droht uns eine Bindestrich Inflation?" *Muttersprache* 103: S. 50–60.

Stede, Manfred (2004): "The Potsdam Commentary Corpus". In: *Proceeding of the ACL-04 Workshop on Discourse Annotation*. Barcelona. `https://aclweb.org/anthology/W/W04/W04-0213.pdf`.

Stede, Manfred (2007): *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*. narr Studienbücher. Tübingen: Gunter Narr.

Steffens, Doris und al Wadi, Doris (2013): *Neuer Wortschatz. Neologismen im Deutschen 2001-2010. 2 Bände*. Berlin: De Gruyter.

Storrer, Angelika (2000): "Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet". In: *Neue Medien im Alltag: Begriffsbestimmungen eines interdisziplinären Forschungsfeldes*, herausgegeben von Voß, Gerd-Günter; Holly, W. und Boehnke, K., Opladen, S. 151–176.

Storrer, Angelika (2001): "Getippte Gespräche oder dialogische Texte? Zur kommunikativen Einordnung der Chat-Kommunikation". In: *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik*, herausgegeben von et al., Andrea Lehr, Berlin / New York, S. 439–465.

Storrer, Angelika (2006a): "Funktionen von nominalisierungsverbgefügen im text. eine korpusbasierte fallstudie". In: *Von der Intentionalität zur Bedeutung konventionalisierter Zeichen. Festschrift für Gisela Harras zum 65. Geburtstag*, herausgegeben von Prost, Kristel und Winkler, Edeltraud, Tübingen: Narr, S. 147–178.

Storrer, Angelika (2006b): "Zum Status der nominalen Komponente in Nominalisierungsverbgefügen". In: *Grammatische Untersuchungen. Analysen und Reflexionen*, herausgegeben von Breindl, Eva; Gunkel, Lutz und Strecker, Bruno, Tübingen: Narr, S. 275–295.

Strube, Michael und Hahn, Udo (1999): "Functional Centering – Grounding Referential Coherence in Information Structures". *Computational Linguistics* 25: S. 309–344. `https://aclweb.org/anthology/J/J99/J99-3001.pdf`.

Stubbs, Michael (1996): *Text and corpus analysis: Computer-assisted studies of language and culture*, Band 23 von *Language in society*. Oxford: Blackwell.

Svartvik, Jan (Herausgeber) (1992): *Directions in corpus linguistics: Proceedings of Nobel symposium 82, Stockholm, 4 - 8 August 1991*, Band 65 von *Trends in Linguistics: Studies and Monographs*. Berlin / New York: Mouton de Gruyter.

Telljohann, Heike; Hinrichs, Erhard W. und Kübler, Sandra (2004): "The tüba-d/z treebank: Annotating german with a context-free backbone". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lissabon.

Telljohann, Heike; Hinrichs, Erhard W. Kübler, Sandra; Zinsmeister, Heike und Beck, Kathrin (2012): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen. `http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1201.pdf`.

Tesnière, Lucien (1959): *Éléments de syntaxe structurale*. Paris: Klincksieck.

Thomas Bartz, Michael Beißwenger, Angelika Storrer (2013): "Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge". *JLCL* 28 (1): S. 157–198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf.

Thurmair, Maria (1989): *Modalpartikeln und ihre Kombinationen*. Band 223 von *Linguistische Arbeiten*. Tübingen: Niemeyer.

Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Band 6 von *Studies in Corpus Linguistics*. Amsterdam: Benjamins.

Tomášiková, Slavomíra (2008): "Okkasionalismen in den deutschen Medien". In: *Médiá a text II*. herausgegeben von Bočák, Michal und Rusnák, Juraj. Prešove, S. 246–256.

Ueberwasser, Simone (2013): "Non-standard data in Swiss text messages with a special focus on dialectal forms". In: *Non-Standard Data Sources in Corpus-Based Research*. herausgegeben von Zampieri, Marcos und Diwersy, Sascha. Aachen: Shaker Verlag. S. 7–24.

Ule, Tylman und Hinrichs, Erhard (2004): "Linguistische Annotation". In: *Texttechnologie – Perspektiven und Anwendungen*. herausgegeben von Lobin, Henning und Lemnitzer, Lothar. Tübingen: Stauffenburg, S. 217–343.

Uszkoreit, Hans; Brants, Thorsten; Duchier, Denys; Krenn, Brigitte; Konieczny, Lars; Oepen, Stephan und Skut, Wojciech (1998): "Studien zur performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen". CLAUS Report 99, Universität des Saarlandes. ww.osgk.ac.at/~brigitte.krenn/papers/rel_satz_springer98.pdf.

Volk, Martin (1995): *Einsatz einer Testsatzsammlung im Grammar Engineering*. Band 30 von *Sprache und Information*. Tübingen: Niemeyer.

Wagner, Andreas und Zeisler, Bettina (2004): "A syntactically annotated corpus of Tibetan". In: *Proceedings of LREC 2004*. Lissabon, S. 1141–1144. http://www.lrec-conf.org/proceedings/lrec2004/pdf/293.pdf.

Walter, Maik (in Vorbereitung): *Der Gebrauch von Konnektoren in fortgeschrittenen Lernervarietäten: Eine korpusbasierte Analyse*. Dissertation, Humboldt-Universität, Berlin.

Weber, Heinz J. (1997): *Dependenzgrammatik. Ein interaktives Arbeitsbuch*. narr studienbücher. Tübingen: Gunter Narr, 2. Auflage.

Wermter, Joachim und Hahn, Udo (2004): "Collocation Extraction Based on Modifiability Statistics". In: *COLING'04 - Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland*. Genf. https://aclweb.org/anthology/C/C04/C04-1141.pdf.

Weydt, Harald (Herausgeber) (1979): *Die Partikeln der deutschen Sprache*. Berlin: De Gruyter.

Weydt, Harald (1983): *Partikeln und Interaktion*. RGL. Tübingen: Niemeyer.

Widdows, Dominic (2004): *Geometry and Meaning*. Stanford: CSLI publications.

Wiegand, Herbert Ernst (1998): *Wörterbuchforschung. Untersuchungen zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie, 1. Teilband*. Berlin / New York: Mouton de Gruyter.

Wittgenstein, Ludwig (1967): *Philosophische Untersuchungen*. Frankfurt/M.: Suhrkamp.

Yang, Dan Hee; Lee, Ik-Hwan und Cantos, Pascual (2002): "On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition". *Computers and the Humanities* 36: S. 171–190.

Zeldes, Amir (erscheint): "The Case for Caseless Prepositional Constructions with *voller* in German". In: *Constructional Approaches to Syntactic Structures in German*, herausgegeben von Boas, Hans C. und Ziem, Alexander, Berlin: De Gruyter.

Zeldes, Amir; Zipser, Florian und Neumann, Arne (2013): "PAULA XML documentation". Forschungsbericht. 2013, 38 Seiten. <hal-00783716> http://hal.inria.fr/hal-00783716/.

Zinsmeister, Heike (2011): "Chancen und Probleme der Nutzung von Korpora, Taggern und anderen Sprachressourcen in Seminaren". *JLCL* 26 (1): S. 67–79. http://media.dwds.de/jlcl/2011_Heft1/6.pdf.

Zinsmeister, Heike (2015): "Chancen und Grenzen von automatischer Annotation". *Maschinelle Textanalyse, Themenheft der Zeitschrift für Germanistische Linguistik (ZGL)* 43 (1): S. 84–111.

Zinsmeister, Heike und Breckle, Margit (2012): "The ALeSKo learner corpus: design – annotation – quantitative analyses". In: *Multilingual Corpora and Multilingual Corpus Analysis*, herausgegeben von Schmidt, Thomas und Wörner, Kai, Amsterdam: John Benjamins, Hamburg Studies in Multilingualism, S. 71–96.

Zinsmeister, Heike und Heid, Ulrich (2003): "Significant Triples: Adjective+Noun+Verb Combinations". In: *Proceedings of Complex*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.404.4254&rep=rep1&type=pdf.

Index

Die **fettgedruckten** Ziffern weisen auf Textstellen hin, an denen der Indexterm definiert wird.

Die linguistische Arbeit mit digitalen Textsammlungen hat sich in den letzten Jahren von einer Methode zu einer eigenen Disziplin der Linguistik entwickelt. Im Zentrum des Buches stehen methodische Fragen, die Darstellung deutschsprachiger Korpora und die Diskussion jüngerer Arbeiten mit korpuslinguistischem Bezug. Die Autoren wenden sich dabei insbesondere an Lehrende und Studierende der Germanistik, die Korpora in ihre eigenen Forschungsarbeiten einbeziehen möchten, und an theoretische Linguisten, die ihre Theorien an authentischen Sprachdaten überprüfen wollen.

Die 3. Auflage diskutiert die Nutzung von Korpora als linguistische Evidenz, führt an ihre quantitative Auswertung heran und enthält neue Fallstudien zur internetbasierten Kommunikation und historischen Texten.

Pressestimmen:
„Während Einführungstexte zu Recht neutral in neue Themengebiete einführen müssen, wird hier zusätzlich die Möglichkeit genutzt, die Korpuslinguistik als ein spannendes und lebendiges Gebiet innerhalb der Sprachwissenschaft zu zeigen."

(Zeitschrift für Sprachwissenschaft 26,2 (2007))

„allen Interessierten, seien es Studierende oder Linguisten, zu empfehlen."
(Info DaF Nr. 2/3, April/Juni 2008)

narr STUDIENBÜCHER

ISBN 978-3-8233-6886-1

narr francke attempto