

Lothar Lemnitzer
Heike Zinsmeister

Korpuslinguistik

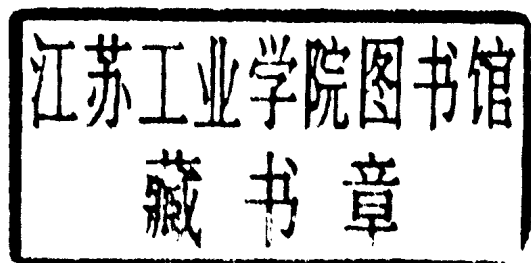
Eine Einführung

narr studienbücher

Lothar Lemnitzer
Heike Zinsmeister

Korpuslinguistik

Eine Einführung



gnV Gunter Narr Verlag Tübingen

Lothar Lemnitzer, Dr. phil., lehrt am Seminar für Sprachwissenschaft der Universität Tübingen und war dort u.a. Koordinator des BA/MA-Studiengangs International Studies in Computational Linguistics; zahlreiche Veröffentlichungen, Autor der »Wortwarte«.

Heike Zinsmeister, Dr. phil., lehrt am Seminar für Sprachwissenschaft der Universität Tübingen und forscht dort im Sonderforschungsbereich »Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung«.

Bibliografische Information der Deutschen Bibliothek

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

© 2006 · Narr Francke Attempto Verlag GmbH + Co. KG
Dischingerweg 5 · D-72070 Tübingen

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.
Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

Internet: <http://www.narr-studienbuecher.de>
E-Mail: info@narr.de

Druck: Gulde, Tübingen
Bindung: Nägele, Nehren
Printed in Germany

ISSN 0941-8105
ISBN 3-8233-6210-0

Inhalt

1	Einleitung	5
1	Was ist Korpuslinguistik?	5
2	Wer sollte dieses Buch lesen?	11
3	Aufbau des Buchs	11
2	Die Quellen linguistischer Erkenntnis	14
1	Empirismus und Rationalismus	15
2	Sprecherurteile statt Korpusdaten – Die Position der Generativen Grammatik	19
3	Linguistische Erkenntnis geht vom Sprachgebrauch aus – Die Position des Kontextualismus	28
4	Korpusbasierte Ansätze	32
5	Weiterführende Literatur	38
6	Aufgaben	38
3	Linguistische Korpora	40
1	Definition und Abgrenzung	40
2	Primärdaten und Metadaten	44
3	Methodische Probleme und ihre Lösung	50
4	Methodisches Vorgehen beim Aufbau eines Korpus – Eine Anleitung	57
5	Weiterführende Literatur	59
6	Aufgaben	59
4	Linguistische Annotation und ihre Nutzung	60
1	Motivation	60
2	Annotationsebenen	63
3	Korpusabfrage	88
4	Annotation Ihres eigenen Korpus	95
5	Weiterführende Literatur	97
6	Aufgaben	98
5	Deutschsprachige Korpora	101
1	Einleitung	101
2	Korpus typologie	102
3	Deutsche Korpuslandschaft	107
4	Weiterführende Literatur	126

6	Korpuslinguistik in der Praxis	127
1	Übersicht	127
2	Orthographie	128
3	Wortbildung	130
4	Syntax	136
5	Computerlinguistik	141
6	Lexikologie und Lexikographie	143
7	Partikeln	157
8	Besondere Textsorten	164
9	Sprachlehre	166
10	Fazit	170
11	Weiterführende Literatur	172
12	Aufgaben	173
7	Erfahrungen von Linguisten mit Korpora	175
1	Porträts	175
2	Fragen und Antworten	177
8	Glossar	196
	Literaturverzeichnis	199
	Index	215

Vorwort

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 (<http://www.lemnitzer.de/lothar/KoLi>).

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 Berlin

Einleitung

1 Was ist Korpuslinguistik?

Die Folklore der Sprachwissenschaft¹ kennt zwei Forschertypen:

Der Denker. Der Denker² 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, deren Existenz durch die Grammatik, die dieser Sprachtheorie entspricht, hergeleitet werden kann. 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. 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

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.

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.

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 Sprachunterricht, geht.

Diese plastische Beschreibung zweier Typen von Forschern in der Linguistik ist nicht neu. Sie findet sich so ähnlich schon bei Charles 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 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 Zukunft 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

³ Vgl. Hockett (1964), zitiert nach McEnery und Wilson (1996). Wir werden in Abschnitt 2 auf die Probleme eingehen, die Selbstauskünfte von Sprechern tatsächlich mit sich bringen.

der Korpuslinguistik als irrelevant und nutzlos⁴. 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* und *Linguistik* zusammen. Eine Antwort auf die Frage führt also zunächst über diese beiden Begriffe.

Definition 1 (Korpus⁶). *Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen⁷. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.*

Die Sammlung von Texten kann zufällig entstanden sein oder als Ergebnis sorgfältiger Planung. 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 nichtdigitalisierte Textsammlungen bzw. Recherchen, die sich auf solche beziehen. Wir werden im Kapitel *Korpuslinguistik in der Praxis* solche Untersuchungen vorstellen. Die Verwendung nichtdigitalisierter Texte führt jedoch zu methodischen Problemen. Auch dies werden wir in Kapitel 6 zeigen. Ältere Texte werden heute in vielen Projekten nachträglich digitalisiert. Das Gleiche gilt für Tonaufzeichnungen von Interviews, Gesprächen u.s.w. Man tut gut daran, sich Gedanken zu machen, ob es digitalisierte Daten für die eigenen

⁴ Z.B. Chomsky (1986), S. 27.

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

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 Wir werden im Folgenden in der Regel von *Texten* reden und nur dort den Ausdruck *Äußerung* verwenden, wo es ausdrücklich um gesprochene Sprache geht.

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 u.s.w. Man spricht hierbei auch von *Metadaten*. Von diesen Daten, die ganze Texte oder zusammenhängende Äußerungsfolgen beschreiben, unterscheiden wir die linguistische Annotation, die sich immer auf Teile von Äußerungen bezieht, also auf Wörter, Sätze usw. Zu diesen linguistisch relevanten Einheiten wird zum Beispiel deren linguistische Kategorie oder grammatische Funktion angegeben.

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* Korpora⁸.

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:

- (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

⁸ Ein multimodales Korpus entstand im Sonderforschungsbereich 441, Linguistische Datenstrukturen, wo für das Bosnische / Serbische / Kroatische unter anderem am Beispiel von Comicbildern und deren Texten lokale und temporale Deiktika (Zeigewörter) untersucht wurden, vgl. <http://www.sfb441.uni-tuebingen.de/b8/> und Raecke (2000).

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 es, 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 konzentrieren⁹. Inwieweit von Texten 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 sollte¹⁰.

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

Definition 2 (Korpuslinguistik). *Als Korpuslinguistik bezeichnet man 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 Tätigkeit, d.h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Korpusbasierte Sprachbeschreibung kann verschiedenen Zwecken dienen, zum Beispiel dem Sprachunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen Sprachverarbeitung.*

⁹ 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.

¹⁰ „... 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.

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*¹¹. Eine Stärke der Korpuslinguistik ist es, dass auf Grund der Datenbasis nicht nur die Struktur einer Sprache, sondern auch deren Verwendungsweise(n) untersucht werden können.

Die Einhaltung gewisser Prinzipien ist die Grundvoraussetzung jeder wissenschaftlicher Tätigkeit. Dazu gehört, dass die Ergebnisse von Untersuchungen nachprüfbar oder sogar reproduzierbar sein müssen. Im Fall der Korpuslinguistik heißt das, dass die Ergebnisse von Untersuchungen an vergleichbaren, anderen Korpora als denen, auf die sie sich stützen, nachprüfbar sein sollten. Die gemeinsame Nutzung eines Korpus für verschiedene Untersuchungen gewährleistet, dass Forschungsergebnisse miteinander verglichen werden können. Die Methoden der Untersuchung sollten den anerkannten wissenschaftlichen Standards entsprechen¹², und es muss Klarheit bestehen über die Reichweite und Sicherheit von Aussagen, die auf Grund von Beobachtungen getroffen werden. Dies trifft gleichermaßen für statistische Aussagen über Regularitäten wie für Gesetzaussagen 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. Gesetzaussagen 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 die Sprachlehre.

¹¹ Vgl. hierzu Altrichter (2001).

¹² Diese müssen für die Korpuslinguistik zum Teil noch entwickelt oder etabliert werden, wozu dieses Buch beitragen möchte.

2 Wer sollte dieses Buch lesen?

Diese Einführung wendet sich an Studierende und Forscher der Sprachwissenschaft, die empirisch die deutsche Sprache untersuchen wollen. Wir möchten 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.

Korpuslinguistik hat, wie wir später noch sehen werden, viel mit den quantitativen Aspekten von Sprache zu tun. Wir werden deshalb nicht umhin kommen, auf die quantitativen Aspekte korpuslinguistischer Forschung einzugehen. Diese werden aber nicht im Mittelpunkt dieser Einführung stehen. 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.

Wir werden lediglich die Kenntnisse voraussetzen, die in einer allgemeinen Einführung in die (germanistische) Linguistik erworben werden können.

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?

Die linguistische Annotation von Korpora ist entscheidend für deren Nutzung in Bezug auf sprachwissenschaftliche Fragestellungen. Manche Fragestellungen lassen sich erst beantworten, wenn die Daten, die herangezogen werden, bereits linguistisch voranalysiert und beschrieben sind. Viele Korpora wurden und werden deshalb mit linguistischen Annotationen versehen. Meist werden die Kategorien der linguistischen Einheiten angegeben, aus denen die Texte des Korpus bestehen. Wir werden im vierten Kapitel Mittel und Methoden der Annotation darstellen. Syntaktisch annotierte Korpora nennt man *Baumbanken*. Wir werden einige Beispiel hierfür im Detail vorstellen. In diesem Kapitel wird außerdem die linguistische Abfrage von Korpora thematisiert. Die wichtigsten heute gebräuchlichen Abfragewerkzeuge werden hier vorgestellt.

Ausgehend von einer Typologie von Korpora werden wir im fünften Kapitel die wichtigsten heute verfügbaren Korpora des Deutschen vorstellen.

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

Im siebten und letzten Kapitel wollen wir Experten zu Wort kommen lassen. Wir haben eine Reihe von Sprachwissenschaftlern, die z.T. schon recht lang mit Korpora arbeiten, interviewt und Ihnen vier Fragen gestellt. Die Antworten sind unter diesen vier Fragen zusammengefasst und regen Sie hoffentlich zu eigenen Reflexionen an.

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 Website. Auf dieser Site, die unter <http://www.lemnitzer.de/lothar/KoLi> erreichbar ist, finden Sie:

-
- Weitere Details zu den im Kapitel *Deutschsprachige Korpora* vorgestellten Korpora;
 - 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.

Wir wünschen Ihnen viel Spaß bei der Arbeit mit diesem Buch!

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 wird. Sie werden die unterschiedlichen erkenntnistheoretischen Positionen, auf denen 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 jüngst 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 Korpusdaten¹ 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.

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.