Einleitung

1 Was ist Korpuslinguistik?

Die Folklore der Sprachwissenschaft[1] kennt zwei Forschertypen:

Der Denker. Der Denker 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 Bild, das er durch diese Beobachtungen gewinnen möchte, sollte möglichst vollständig sein, deshalb ist er vor allem an den Phä-

[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.

nomenen 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 (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 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*)[2]. Viele theoretische Sprachwissenschaftler im Umfeld der generativen Sprachtheorie, allen voran Noam Chomsky, bezeichnen das Werk

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

der Korpuslinguistik als irrelevant und nutzlos[3]. Es gibt, wie gesagt, Berichte von „Lagerwechseln“[4], 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 (Korpus[5]). *Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen[6]. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selber 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 in Kaptitel **??** solche Untersuchungen vorstellen. Die Verwendung nichtdigitalisierter Texte führt jedoch zu methodischen Problemen. Auch dies werden wir in Kapitel **??** zeigen. Ältere Texte werden heute in vielen Projekten nachträglich digitalisiert. Das Gleiche gilt für Tonaufzeichnungen von Interviews,

[3] z.B. Chomsky (1986), S. 27.

[4] vgl. zum Beispiel Fillmore (1992) und Sampson (1996).

[5] 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.

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

Gesprächen u.s.w. 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 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* Korpora[7].

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) Am Abend zuvor werden bei der Eröffnung ... Gedichte in türkischer und deutscher Sprache zum Thema gelesen. (taz)

(2) Der erste Blick von der Bühne ins gut gefüllte Auditorium ... ver-schlug mir die Sprache. (taz)

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

(4) Gegen seine ... an irische Folklore erinnernden Songs scheint sich die Sprache Hölderlins zu wehren. (taz)

In Beispiel (1) ist mit *Sprache* eine konkrete natürliche Sprache, zum Beispiel das Deutsche oder das Türkische, gemeint. In Beispiel (2) geht

[7] 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).

es allgemeiner um das Sprachvermögen und den Zugang zu diesem, welcher bei dem entgeisterten Schauspieler 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 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 konzentrieren[8]. 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 sollte[9].

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 Tätig-*

[8] 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.

[9] „. . . 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

keit, 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.

Gegenstand von Korpora und damit der Korpuslinguistik sind natürliche Sprachen, nicht formale Sprache 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*[10]. 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 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, reproduzierbar sein sollten. Die gemeinsame Nutzung von Korpora 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 Gesetzesausagen 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 beeinflus-

[10] Vgl. hierzu Altrichter (2001).

sen u.a. die Übersetzungswissenschaft, die Lexikografie und die Sprachlehre.

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.

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 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 in Kapitel vier beschriebenen Korpora
- Hinweise auf Werkzeuge, die die Arbeit mit Korpora erleichtern
- Handreichungen zu einigen der gebräuchlicheren Korpuswerkzeuge
- Lösungsänsätze für die Übungsaufgaben
- Weitere nützliche Links
- Weitere Informationen zu den Autoren des Buchs

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

Literaturverzeichnis

Altrichter, Helmut (2001): "Retrodigitalisierung in Deutschland – Versuch einer Zwischenbilanz". `http://www.bsb-muenchen.de/mdz/forum/altrichter/`.

Chomsky, Noam (1986): *Knowledge of Language*. Convergence. New York / Westport / London: Praeger.

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.

Hockett, Charles F. (1964): "Sound Change". *Language* 41: S. 185–204.

McEnery, Tony und Wilson, Andrew (1996): *Corpus Linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh: Edinburgh University Press.

Pullum, Geoffrey K. (1991): *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language*. Chicago: The University of Chicago Press.

Raecke, Jochen (2000): "Zeigen im Comic und Zeigen im Film - oder: Deiktika auf der Schnittstelle von Visualität und Verbalität". In: *Slavistische Linguistik*, herausgegeben von Breu, W., München, S. 161–185.

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.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.