CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus
Daten – Korpora – Dokumentation
huius textus situs retis mundialis
http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/daten.pdf
dies manuscripti postremum modificati
03.04.2006
occasio orationis habitae
42. Jahrestagung des Instituts für Deutsche Sprache Mannheim, 14.–16.03.2006
volumen publicationem continens
Kallmeyer, Werner & Zifonun, Gisela (eds.), *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Berlin & New York: W. de Gruyter (Jahrbuch des Instituts für Deutsche Sprache)
annus publicationis
2007
paginae

Daten – Korpora – Dokumentation

Christian Lehmann

Universität Erfurt

Zusammenfassung

Der Begriff und die Rolle von Daten in einer Wissenschaft hängen eng mit ihrem Selbstverständnis zusammen. Als erstes ist zu überlegen, inwiefern Linguistik eine empirische Wissenschaft ist und also von Daten abhängt.

Während in den Philologien ein Korpus die Grundlage einer Disziplin abgibt, die ohne es nicht bestünde, ist in der Linguistik ein Korpus nur ein Weg, an Daten zu kommen. Hier ist zu diskutieren, welche relativen Meriten die alternativen Wege im Hinblick auf die angestrebten Ziele haben.

Während manches auf uns gekommene Korpus seine Sprache sicher nicht angemessen repräsentiert, könnte eine heute von Linguisten erstellte Dokumentation diesen Anspruch im Prinzip einlösen. Hier stellt sich die Frage, ob das – angesichts des infiniten Charakters der Sprache – überhaupt möglich ist und in wieweit die Repräsentativität wieder von den angestrebten Zielen abhängt.

Dies sind alles Fragen linguistischer Methodologie. Eine Zeitlang hat man in der Linguistik geglaubt, ohne Methodologie zu Theorien gelangen zu können. Seit sich das als irrig herausgestellt hat, ist die Entwicklung von Methoden ein fühlbares Desiderat geworden. Wie man repräsentative Daten erhebt, wie man ein Korpus zusammenstellt und nutzt, wie man eine Sprache dokumentiert, sind alles Fragen, die eigentlich in die Alltagsroutine einer Wissenschaft fallen müssten. Dass sie noch weitgehend ungeklärt sind, ist ein Symptom dafür, dass die Linguistik noch keine erwachsene Wissenschaft ist.[1]

1 Einleitung

Das Thema dieser Tagung und dieses Beitrags ist in erster Linie ein methodologisches. Die Fragen nach der Natur von Daten, von Korpora und von Dokumentation, ihrer Rolle in unserer Wissenschaft und in der Praxis kommen zwar je für sich auch in anderen Disziplinen vor; in ihrer Kombination aber sind sie typisch für die Linguistik. Sie hängen eng mit der Frage zusammen, was für eine Wissenschaft Linguistik eigentlich ist, und scheinen deren Klärung vorauszusetzen. Tatsächlich ist dies aber kein einseitiges Voraussetzungsverhältnis, weil Entscheidungen über den Gegenstand und die Methodologie einer Wissenschaft ihr Wesen prägen.

[1] Ich danke Ulrich Ammon, Hans Uszkoreit und Herbert-Ernst Wiegand für hilfreiche Kommentare.

Die obersten Entscheidungen, die das Wesen einer Wissenschaft bestimmen, sind die folgenden (vgl. Abb. 1):

1. Da eine Wissenschaft eine menschliche Aktivität ist, ist sie in erster Linie durch ihr Ziel bestimmt. Da Wissenschaft das Streben nach objektiver Erkenntnis ist, ist jede einzelne Wissenschaft durch ihr spezifisches **Erkenntnisinteresse** bestimmt.
2. Das Erkenntnisinteresse konstituiert den Gegenstand einer Wissenschaft. Eine erste binäre Einteilung folgt dem Kriterium, ob sich der Gegenstand außerhalb des erkennenden Geistes befindet oder nicht. Im letzteren Falle ist es **logische Erkenntnis**. Im ersteren Falle unterscheidet das weitere Kriterium, ob der Gegenstand Produkte eines (ebenfalls erkennenden) Geistes sind oder nicht, zwischen **hermeneutischer** und **empirischer Erkenntnis**.

Abb. 1. *Arten der Erkenntnis*

Erkenntnisobjekt	↙ *im erkennenden Geist*	↘ *außerhalb des erkennenden Geistes*	
		↙ *Produkt eines anderen erkennenden Geistes*	↘ *natürlicher Gegenstand*
	↓	↓	↓
Erkenntnisart	logisch	hermeneutisch	empirisch

Viele Disziplinen lassen sich nach diesen Kriterien zwanglos kategorisieren. Die Linguistik scheint hiernach eindeutig eine hermeneutische Wissenschaft zu sein. Das war auch durch den größten Teil ihrer Geschichte, vom Altertum bis in die Neuzeit, überwiegend so. Während der Neuzeit bildete sich jedoch für die Wissenschaften das Ideal **objektiver Erkenntnis** aus; und intern werden wissenschaftliche Aktivitäten danach bewertet, wie nahe sie diesem Ideal kommen.

Objektive Erkenntnis ist nun

- im Hinblick auf das erkennende Subjekt intersubjektiv
- und im Hinblick auf das Erkenntnisobjekt verallgemeinerbar.

In vollständig abgesicherter Weise wird dieses doppelseitige Objektivitätserfordernis nur in logischer Erkenntnis eingelöst, aber eben um den Preis, mit nichts zu tun zu haben, was sich außerhalb des erkennenden Geistes befände. Ziemlich erfolgreich in dieser Hinsicht ist auch noch die empirische Erkenntnis. Sie sieht von partikulären Eigenschaften ihres Gegenstandes ab und erreicht allgemeine Aussagen in Form von Hypothesen. An der Intersubjektivität muss sie Abstriche machen, denn die Hypothesen sind nicht verifizierbar, sondern nur falsifizierbar. Hermeneutische Erkenntnis schließlich ist in bezug auf beide Pole der Erkenntnis partikulär, insofern sie vom erkennenden Subjekt abhängt und insofern sie sich auf ein individuelles Objekt bezieht.

Während die jüngeren Fortschritte der Hermeneutik in Richtung einer objektiven Hermeneutik (vgl. z.B. Oevermann 2002) unbestritten sind, ergibt das Kriterium der Objektivität doch eine Hierarchie der Wissenschaftlichkeit unter den bestehenden Disziplinen:

Ein unangefochtenes Prestige von Wissenschaftlichkeit haben nur die logischen Disziplinen. Die empirischen Disziplinen werden seit Hempel 1966, soweit eben möglich, nach dem Modell der logischen Disziplinen aufgezogen – z.B. die Physik nach dem Modell der Mathematik. Die hermeneutischen Disziplinen hinwiederum gelten überhaupt nicht als *sciences*, sondern lediglich als *humanities*. Manche von ihnen, darunter die Psychologie, die Soziologie und die Linguistik, wollen am liebsten keine hermeneutischen Disziplinen sein. Sie behandeln die Unterscheidung zwischen Erzeugnissen des Mitmenschen und anderen erfahrbaren Gegenständen als irrelevant und gehen nach Möglichkeit empirisch oder gar logisch vor.

Im Falle der Linguistik ist dies bis zu einem gewissen Grade durch die Natur des Gegenstandes gerechtfertigt. Sprachzeichen haben eine wahrnehmbare Seite, Sprechakte sind beobachtbare raumzeitliche Ereignisse. Man kann sie empirisch angehen. Ferner ist Sprache, in ihrer kognitiven Dimension, die äußere Form des Denkens und insofern logischer Erkenntnis zugänglich. Beides ändert freilich nichts daran, dass Sprachtätigkeit, in ihrer sozialen Dimension, Verständigung ist und dass diese hermeneutischer Methodik bedarf. Der Gegenstand Sprache hat wesentlich diese drei Aspekte, den hermeneutischen, den empirischen und den logischen. Jeglicher Versuch, einen davon zu eskamotieren oder zu verabsolutieren, ist steriler Reduktionismus.[2]

Dies vorausgeschickt, wird die Linguistik im folgenden überwiegend in ihrer Eigenschaft als empirische Disziplin behandelt. Dies ist einfach eine Folge der Themenstellung: Daten gibt es nur in empirischen Disziplinen.

2 Daten

2.1 Funktion von Daten

Ein **Datum** ist eine Repräsentation eines Phänomens aus dem Gegenstandsbereich einer Wissenschaft, die als gegeben angenommen wird (vgl. Lehmann 2004). Die Repräsentationsbeziehung ist nicht frei manipulierbar, sondern durch die wissenschaftliche Methodik geregelt. Das ist eben die Basis dafür, dass das Datum als gegeben angenommen wird. Es vertritt für die Zwecke wissenschaftlicher Argumentation das Phänomen selbst, welches i.a. nicht zur Hand ist. Daten spielen in einer Disziplin eine methodische Rolle in dem Maße, in dem sie eine empirische Disziplin ist. In induktiver Methodik dient das Datum als Indiz und empirische Evidenz,[3] in deduktiver Richtung als Prüfstein im Test einer Theorie.

Eine Wissenschaft konstituiert sich gemäß ihrem **Erkenntnisinteresse**. Es gibt legitime oder jedenfalls in der Gesellschaft unangefochtene Erkenntnisinteressen, die nichts mit in der uns umgebenden Welt vorfindlichen und öffentlich beobachtbaren Phänomenen zu tun haben. Für die Linguistik aber gilt in dieser Hinsicht folgendes:

---

[2] Die drei Aspekte sind in linguistischer Forschung miteinander verwoben. Aber in gewissen Forschungsrichtungen dominiert einer von ihnen. Z.B. ist formale Semantik überwiegend logische, statistische Korpuslinguistik überwiegend empirische und Konversationsanalyse überwiegend hermeneutische Wissenschaft.

[3] Die Forderung, die Daten, auf welche Generalisierungen sich stützen, verfügbar zu machen, wird immer häufiger erhoben. S. z.B. Corbett 2005, §8 über "Reproduzierbarkeit" von typologischen Ergebnissen dadurch, dass die Daten, auf denen sie basieren, zur Verfügung gestellt werden.

1. Eine Disziplin, die sich für eine empirische erklärt, muss ihren Gegenstandsbereich mit Bezug auf unabhängig von ihr selbst in der Welt vorfindliche und öffentlich beobachtbare Phänomene konstituieren. In dem Maße, in dem sie das nicht tut, ist es vielleicht eine logische oder hermeneutische Disziplin oder vielleicht auch Esoterik.
2. Es gibt in der uns umgebenden Welt vorfindliche und öffentlich beobachtbare Phänomene, nämlich Sprechakte im weitesten Sinne, an deren wissenschaftlicher Erforschung ein öffentliches Interesse und geradezu ein Auftrag besteht und für die keine andere Disziplin zuständig ist.
3. Die Linguistik selbst hat sich, in Gestalt herausragender Vertreter wie Leonard Bloomfield und Noam Chomsky, für eine empirische Disziplin (und gelegentlich gar, in rhetorischer Polarisierung, für eine Naturwissenschaft) erklärt. Das erfordert, dass sie ihren Gegenstandsbereich auf die in Nr. 1 beschriebene Weise konstituiert. Dessen vorgegebene Basis sind gerade die in Nr. 2 genannten Phänomene.
4. Jegliche Wissenschaft, die über reine Idiographie hinausgelangen will, arbeitet mit Abstraktionen. Das gilt jedenfalls für die Linguistik und a fortiori für die Systemlinguistik und ihren Gegenstand, das Sprachsystem. Von den notwendigen und angestrebten Abstraktionen sind zu allererst die Daten betroffen. Sie werden gewisser Eigenschaften entkleidet, denen das Erkenntnisinteresse nicht gilt. So entstehen sekundäre Daten, die oft eine sehr vermittelte Beziehung zu den Phänomenen haben.
5. Wie jegliche empirische Wissenschaft kann auch die Linguistik ihren Anspruch, eine solche zu sein, nur aufrecht erhalten, wenn sie die Beziehung ihrer Theorien und a fortiori ihrer Daten auf die Phänomene jederzeit methodisch kontrolliert und objektiv nachvollziehbar macht. In dem Maße, in dem sie das nicht tut, ist ihr Anspruch, eine empirische Disziplin zu sein, wissenschaftspolitische Rhetorik.
6. Gemäß dem in § 1 Gesagten ist die Linguistik sowohl eine empirische als auch eine logische als auch eine hermeneutische Disziplin. Sie hängt also nicht ausschließlich von Daten ab, sondern gewinnt Erkenntnis auch auf andere Weise. Desto komplexer ist ihre Methodologie und desto größer die Gefahr der Manipulation von Ergebnissen.

Die Frage, welcher Art die Daten sind und welche Rolle sie in der Methodologie spielen, berührt das Selbstverständnis einer Disziplin auf das Intimste. Die Linguistik bestand jahrtausendelang nur in Form von logischen und hermeneutischen Disziplinen und wird erst seit dem 20. Jh. ernstlich als empirische Wissenschaft aufgefasst.[4] Der Begriff des sprachlichen Datums ist deshalb noch unterentwickelt. Bis in die Gegenwart haben Linguisten in dem Bemühen, Evidenz für ihre Hypothesen anzuführen, immer wieder Beispielsätze als Daten ausgegeben. Nun haben **das Datum und das Beispiel** in der Tat so viel gemeinsam, dass beide Repräsentationen von Phänomenen des Gegenstandsbereichs sind. Aber sie haben völlig verschiedene Funktion und daher verschiedenen methodologischen Status. Ein Beispiel dient der Veranschaulichung eines Theorems. Es hat einen kommunikativen und oft gar didaktischen Zweck, denn es soll beim Leser Verständnis sichern. Dazu muss es solche Eigenschaften aufweisen, auf die es im gegebenen Zusammenhang ankommt; und andere Eigenschaften, die ablenken oder irreführen könnten, sollte es nicht aufweisen. Daten, die diese Anforderungen erfüllen, sind oft schwer oder überhaupt nicht zu finden. Daher ist es üblich, dass der Autor des Theorems auch das Beispiel selbst bildet. Dagegen ist nichts

[4] Die Idee, die Linguistik sei eine Naturwissenschaft, findet sich zwar bereits im 19. Jh. Aber da überwog die wissenschaftspolitische Rhetorik wohl die Bereitschaft, ernstlich auf der Basis von Primärdaten empirisch zu arbeiten.

einzuwenden, solange niemand glaubt, das Beispiel sei ein Datum. Als solches müsste es nämlich vom Wissenschaftler unabhängig sein, sonst besteht keine Veranlassung, es als gegeben anzunehmen. Folglich können Daten ggf. als Beispiele dienen, aber nicht umgekehrt.

Weite Teile der Linguistik des 19. und 20. Jh. sind, was die Methodologie und insbesondere die Daten angeht, nahtlose Fortsetzungen jahrtausendealter Vorgeschichte. Die philosophische Grammatik des Altertums, z.B. Aristoteles, arbeitete ausschließlich mit ausgedachten Beispielsätzen. Dieser Brauch wurde von der Schulgrammatik übernommen. Der europäische Strukturalismus, z.B. F. de Saussure, folgte demselben Usus. Im amerikanischen Strukturalismus verfuhren Bloomfield und dann insbesondere Chomsky so.

Tatsächlich vorgekommene Äußerungen als Grundlage empirischer Linguistik haben natürlich ebenfalls eine Vorgeschichte. Da ist zunächst die Philologie, deren Aufgabe es ist, vorhandene Texte zu verstehen. In ihrem Gefolge sind z.B. die Beispielsätze, mit denen die historisch-vergleichende Sprachwissenschaft des 19. Jh. arbeitet, überwiegend aus dem überlieferten Korpus entnommen. In den U.S.A. begründet ab 1900 Franz Boas eine Tradition, die eine Sprachbeschreibung aus einem Korpus ableitet. Der Übergang von Korpora schriftlicher zu Korpora mündlicher Sprache ist dann einerseits eine Frage des technischen Fortschritts, andererseits aber auch eine Frage der linguistischen Methodik, die es erlaubt, Daten gesprochener Sprache zu repräsentieren und zu analysieren und die erst ausgearbeitet werden musste.

2.2 Abstraktionen aus Daten

Die Systemlinguistik interessiert sich für Sätze und andere sprachliche Konstruktionen derselben Art. Sätze kommen in der Wirklichkeit nicht vor. Man kann sie induktiv als Abstraktionen über Äußerungen gewinnen. Dazu sieht man von den raumzeitlichen Koordinaten der Äußerung und somit von der Situation, in welcher sie verwendet wurde, ab und fokussiert stattdessen auf ihre rein sprachliche Struktur. Sätze sind somit dekontextualisiert mindestens in dem Sinne, wo der außersprachliche Kontext gemeint ist, meist aber außerdem dekontextualisiert in bezug auf sprachlichen Kontext. Das bedeutet natürlich, dass der Sinn, in dem die Äußerung tatsächlich verwendet wurde, verloren geht. Daraus folgt übrigens die vielleicht wichtigste Beschränkung der Systemlinguistik: Ihr Gegenstand ist zwar letztlich die Verständigung; aber der Sinn, welcher in der Verständigung konstruiert wird, bleibt ihr unzugänglich.

Ein Satz steht also um eine Abstraktionsstufe höher als eine Äußerung. Er gewinnt seinen empirischen Status ausschließlich daraus, dass er als Äußerung verwendbar ist. Dies wird ausschließlich dadurch überprüft, dass er de facto als Äußerung verwendet wird. Methodisch kann man hier im Prinzip in zwei Richtungen vorgehen:

1. Man kann zunächst einen Satz bilden, diesen dann als Äußerung verwenden und überprüfen, ob die Äußerung erfolgreich war.
2. Man kann vorhandene erfolgreiche Äußerungen aufnehmen und durch methodisch kontrollierte Abstraktion in Sätze überführen.

Den beiden Methoden ist das Problem gemeinsam, dass man Kriterien des Erfolgs von Äußerungen braucht. Die erste Methode hat das zusätzliche Problem der Praktikabilität. Insofern ist es wesentlich sinnvoller, sich als Systemlinguist mit der Analyse solcher Sätze zu befassen, die ihren empirischen Status als Äußerung bereits unter Beweis gestellt haben –

eben mit aus Korpora destillierten Sätzen – als zu versuchen, selbst gebildete Sätze als Äußerungen zu lancieren.

2.3 Kompetenzdaten und Performanzdaten

In den allerletzten Jahrzehnten haben verschiedene Richtungen der Linguistik den postulierten empirischen Status dieser Wissenschaft so interpretiert, dass die Datenerhebung endlich ernst genommen werden müsste. Die erste Strömung dieser Art war in den 1970er Jahren die Variationslinguistik. Es folgten verschiedene funktionalistische Richtungen (vgl. etwa Givón 1984:8) und schließlich die Korpuslinguistik. Mehrere der Postulate dieser Strömungen waren explizit gegen die jahrzehntelange Übung gerichtet, Datenbeschaffung durch Introspektion zu ersetzen.

In dem Maße, in dem diese Attacken von verschiedenen Seiten Wirkung zeitigten, gerieten die sterilsten Varianten der Systemlinguistik in die Defensive. Die Verteidigungsstrategie ist seit einiger Zeit eine Unterscheidung zwischen Kompetenz- und Performanzdaten.[5] Sie setzt die Unterscheidung von **Kompetenz und Performanz** voraus, die in Chomsky 1965:4 eingeführt und seitdem nie wesentlich revidiert wurde. Die Unterscheidung hat bei Chomsky eine zentrale theoretische Funktion. Außerhalb seiner Linguistik wird die Unterscheidung höchstens als „a methodologically useful preliminary" (Givón 1984:8) akzeptiert. Insofern die Unterscheidung außerhalb der generativen Linguistik keinen theoretisch relevanten Status hat, entfaltet auch die Verteidigungsstrategie ihre Wirkung nur innerhalb des Modells.

So etwas wie Kompetenzdaten gibt es nicht. „Kompetenzdaten" ist nur ein neuer Name für durch Introspektion gewonnene Beispielsätze und deskriptive Aussagen. Solche sind aber, wie gesagt, keine Daten im Sinne einer empirischen Wissenschaft. Kompetenzdaten im Wortsinne kann es auch gar nicht geben; der Ausdruck ist eine contradictio in adiecto. Denn die Kompetenz ist definiert als ein „tacit knowledge", das nicht bewusst gemacht werden kann. Darüber kann es also per definitionem keine Daten geben.

Sinnvollerweise kann man dagegen in diesem Zusammenhang zwei Unterscheidungen machen:

1. Sätze vs. Äußerungen;
2. sprachnutzende vs. sprachreflexive Daten.

Was Unterscheidung 1 betrifft, so können Eigenschaften von Sätzen durch phonologische, grammatische und semantische Analyse, also mit Methoden der Systemlinguistik, herausgebracht werden. Eigenschaften von Äußerungen sind Phänomene wie Lautstärke, Reaktionszeiten oder Häufigkeiten im Korpus. Sie lassen sich mit den Methoden anderer Subdisziplinen, etwa der Phonetik, Psycho- oder Soziolinguistik, messen.

Was Unterscheidung 2 betrifft, so enthalten Korpora sowohl sprachreflexive als auch sprachnutzende Äußerungen, also solche, die von Aspekten von Sprache handeln und solchen, die das nicht tun. Hier kann man empirisch feststellen,

- was die Sprecher auf sprachnutzender Ebene tatsächlich tun,

---

[5] Die Begriffe stammen allerdings nicht aus der Linguistik, sondern der Professionalisierungsforschung. Obwohl sie erst vor wenigen Jahren in der Linguistik aufgetaucht sind, werden sie bereits uneinheitlich verwendet. Mit ‚Kompetenzdaten' meinen die einen (z.B. Krenn) solche Information in elektronischen Korpora, die die Analysten durch Annotation dazugetan haben. Die anderen (z.B. Uszkoreit) meinen vom Linguisten ersonnene Beispiele (im Gegensatz zu vorgefundenen Daten).

- was sie auf sprachreflexiver Ebene darüber meinen,
- und wie das tatsächliche Sprachhandeln und die Ansichten darüber aufeinander bezogen sind.

Auf Kompetenz vs. Performanz lassen sich diese beiden Unterscheidungen aber nicht zurückführen. Wenn wir zum Zwecke des Arguments die erstere Unterscheidung einmal akzeptieren, so gilt jedenfalls, dass jegliches linguistische Datum auf eine Äußerung, also auf Performanz zurückgeht.

3 Korpus

3.1 Begriffsklärung

Ein **Textkorpus** ist eine Menge von Texten, insofern sie in einer wissenschaftlichen Untersuchung genutzt wird. Der Begriff hat sich in den letzten Jahrzehnten in der Linguistik in verschiedener Hinsicht gewandelt. Traditionell verstand man unter einem Korpus die Gesamtheit von Texten einer bestimmten Kategorie, z.B. das Korpus der hethitischen Texte[6] oder das Korpus der Schriften Platons. Für diesen Begriff spielt es noch keine Rolle, ob diese Menge als geschlossene Sammlung verfügbar ist. Dies ist erst eine linguistische Einengung des Begriffs. Und eine jüngere Einengung verlangt, dass ein Korpus in elektronischer Form vorliege.[7]

Das Erfordernis, dass ein Korpus als geschlossene Sammlung vorliege, brachte die Konnotation mit sich, dass das Korpus kompiliert wird. Damit fällt andererseits das Erfordernis der Exhaustivität. Korpus von Texten der Kategorie X heißt nun auch eine Sammlung, die nur eine ausgewählte Teilmenge von X umfasst.[8]

Eine weitere Ausweitung des Begriffs resultiert aus dem Verzicht auf das Erfordernis, dass die größten Untereinheiten eines Korpus Texte sein müssen. Es kann jetzt sprachliches Material jeglicher Art sein; man spricht auch von Korpora von Sätzen, von Informantenurteilen oder von Lexikoneinträgen.

Schließlich war es ein – wenn auch implizites – Korollar des herkömmlichen Korpusbegriffs, dass die enthaltenen Texte unabhängig vom Kompilator zustande gekommen waren. Der Kompilator stellte die Texte nicht her; er sammelte sie bloß. Zeitgenössische Korpora dagegen umfassen auch solches sprachliche Material, welches eigens für das Korpus hergestellt wurde.

Ich regele hier den Sprachgebrauch so, dass ein Korpus eine geschlossen verfügbare Sammlung von Texten einer bestimmten Kategorie ist. Die Erfordernisse, dass es Texte sein und dass diese als geschlossene Sammlung vorliegen müssen, werden beibehalten; die Erfordernisse der Exhaustivität des Korpus und der vorgängigen Existenz der Texte werden

---

[6] In diesem Sinne spricht man auch von einer Korpussprache.

[7] Eine andere Einengung, die sich in diversen zeitgenössischen Definitionen findet, ist, dass ein Korpus für linguistische Zwecke zusammengestellt sei. Diese Bedingung dürfte auf disziplinäre Beschränktheit ihrer Autoren zurückgehen. Korpora wurden und werden in erster Linie von Philologen genutzt, weiter von Juristen und vielen anderen, darunter auch Linguisten.

[8] Hieraus entsteht dann das Problem der Beziehung des Korpus zur Grundgesamtheit. Ein Weg, es anzugehen, ist, das Korpus als Stichprobe im statistischen Sinne zu betrachten. Ein anderer Ansatz wird in § 4.3 gestreift.

fallengelassen. Was durch die Bedingungen ausgeschlossen wird, mag dann unter den Oberbegriff "linguistische Datensammlung" fallen.

Natürlich erbte der herkömmliche Korpusbegriff ein weiteres implizites Korollar vom Textbegriff, nämlich dass es schriftliche Texte sein müssten. Ich nehme an, dass diese Festlegung mit dem Wesen der Begriffe 'Text' und 'Korpus' nie etwas zu tun hatte, sondern lediglich durch den Stand der Technik bedingt war. Immerhin wirft die Tatsache, dass auch beim heutigen Stand der Technik Texte nur dann wissenschaftlich bearbeitet werden können, wenn sie schriftlich repräsentiert sind, interessante methodische Probleme auf. Mindestens so viel muss klar sein, dass ein Textkorpus jedenfalls kein Ausschnitt aus dem Phänomenbereich, sondern eine Repräsentation davon ist. Es enthält einerseits Abstraktionen über den Rohdaten und andererseits Interpretationen des Kompilators. Es ist essentiell, dass diese methodisch kontrolliert sind. Sonst kann man sich die Arbeit mit Korpora gleich schenken; sie würde uns der Empirie keinen Schritt näher bringen.

3.2 Korpusanalyse vs. Lehnstuhllinguistik

Die traditionelle Grammatik schreibt z.T. seit Jahrhunderten Regeln über syntaktische Konstruktionen und den Gebrauch morphologischer Kategorien fort. Untersuchungen des tatsächlichen Gebrauchs anhand von Korpora geben bisweilen überraschende Resultate.

3.2.1 Hauptkonstituentenstellung

Die Idee, dass der einfache Aussagesatz aus nominalem Subjekt und Prädikat bestehe und dass eine typische Konstituenz des Prädikats ein transitives Verb mit einem nominalen direkten Objekt sei, ist in der Linguistik seit langem wohl etabliert. Beispielsätze des Typs von Sapirs *the farmer kills the duckling* finden sich schon in Grammatiken des 19. Jh. Greenbergs (1963) Grundwortstellungstypologie fußt ganz wesentlich auf Sätzen dieser Struktur. Auf sie sind wiederum weitreichende syntaktische Theorien gegründet worden.

Das Problem mit dieser Satzstruktur ist nur, dass sie in den Texten der meisten Sprachen höchst selten ist und in vielen Sprachen überhaupt nicht vorkommt. Auswertungen von Korpora haben ergeben, dass es pro einfachem Satz normalerweise höchstens ein lexikalisch-nominales Satzglied gibt. Der Satz mit lexikalisch besetztem Subjekt und direktem Objekt fristet sein Dasein hauptsächlich in Grammatikbüchern, und die empirische Basis der von ihm handelnden Grundwortstellungstypologie wird somit teilweise fragwürdig. Andererseits zeitigt eine Korpusuntersuchung der Konstituentenstruktur von Verbalsätzen neue Erkenntnisse, z.B. über den Zusammenhang zwischen nominaler und pronominaler Vertretung von Subjekt und direktem Objekt einerseits und den Funktionen von Topic und Fokus in der Informationsstruktur andererseits (vgl. Du Bois 1987, Lambrecht 1994, Kap. 4.5.2). Konstruktionen wie Linksversetzung und Spaltsatz werden neu verstanden: sie dienen nicht nur der Hervorhebung von Topic bzw. Fokus, sondern sie gewährleisten auch die Einhaltung des Prinzips, dass es pro einfachem Satz höchstens ein nominales Satzglied gibt.[9]

[9] Dies ist seinerseits eine Konsequenz der 'one new idea'-Beschränkung (Chafe 1992:92-95).

3.2.2 Das französische Futur

Im Französischen hat man zum Ausdruck des Futurs die Wahl zwischen drei Konjugationskategorien, dem periphrastischen Futur, illustriert in B1, dem synthetischen Futur (B2) und dem Präsens (B3). Die Beispiele stammen aus einem Korpus des gesprochenen kanadischen Französisch, dem *Corpus du français parlé à Ottawa-Hull*.[10]

B1. Bien demain, tu vas aller au Bingo, tu vas gagner.
"Morgen wirst du zum Bingo gehen und gewinnen."

B2. J'ai dit, "Laisse faire, on ira a messe demain matin."
"Ich habe gesagt: 'Lass gut sein, wir gehen morgen früh zur Messe.'"

B3. Il dit, "J'y vas demain matin chez vous".
"Er sagt: 'Ich komme morgen früh zu euch.'"

Was die Verteilung des synthetischen und des periphrastischen Futurs betrifft, so lehrt die traditionelle französische Grammatik, ersteres sei die Default-Variante, während letzteres modale Nuancen oder größere Proximität zum Ausdruck bringe. Bereits die Beispielsätze bestätigen das nicht. Eine Frequenzanalyse des Korpus bringt folgendes Ergebnis: Zukunftsbezug wird durch die drei Konjugationskategorien mit folgender Häufigkeit ausgedrückt:

Abb. 2. *Verteilung der Varianten zum Ausdruck der Zukunft*

Kategorie	**N**	**%**
periphrast. Futur	2.627	73
synthet. Futur	725	20
Präsens	242	7
Summe	3.594	100

Rein quantitativ betrachtet, ist also das periphrastische Futur die Default-Variante. Bei der Untersuchung der Faktoren, die das synthetische Futur begünstigen, stellt sich heraus, dass keiner der von den Grammatikern für ausschlaggebend gehaltenen Faktoren einen statistisch nachweisbaren Einfluss hat. Statt dessen ist der Kontextfaktor, der statistisch betrachtet mit Abstand der relevanteste ist, der durch B4 illustrierte:

B4. Dire que dans quatre cents ans d'ici il va avoir encore des Asselin, puis ils vont encore parler français. Qu'ils parleront pas l'anglais.
"Zu denken, dass es in 400 Jahren immer noch Asselins geben wird und dass sie immer noch Französisch sprechen werden. Dass sie nicht Englisch sprechen werden."

In negativen Sätzen wird fast ausnahmslos das synthetische Futur verwendet. Keine deskriptive oder präskriptive Grammatik weiß etwas davon. Es ist durch diese Korpusanalyse festgestellt worden. Nun bleibt es der deskriptiven Grammatik vorbehalten, diesen Zusammenhang zu klären.

[10] Das folgende ist, einschließlich der Beispiele, ein Referat von Poplack 2001:415-418.

3.2.3 Der deutsche Konjunktiv II

In deutlicher Anlehnung an die lateinische Grammatik lehrte die deutsche Grammatik seit zwei Jahrhunderten, der Konjunktiv II drücke irreale Modalität aus; und zwar drücke der Konjunktiv Imperfekt (*käme*) den Irrealis der Gegenwart, der Konjunktiv Plusquamperfekt (*wäre gekommen*) den Irrealis der Vergangenheit aus. Zu dieser Regel gestand der Duden (1984:158f) in einer Fußnote einen ausnahmsweisen Zukunftsbezug des Konjunktiv Plusquamperfekt zu.

Eine Untersuchung an Texten (Leirbukt 1991) ergab dagegen, dass Beispiele wie B5 (*Neue Zürcher Zeitung* 18.08.1984) für den tatsächlichen Sprachgebrauch repräsentativ sind:

B5. [Kontext: nächsten Monat steht eine Volksabstimmung über den radikaleren von zwei Vorschlägen an.] Ein Baustopp nach Leibstadt allein hätte vermutlich in der Volksabstimmung keine schlechten Chancen gehabt. Hingegen wäre es sonderbar, wenn eine Mehrheit der Stimmenden dafür gewonnen werden könnte, die fünf bestehenden Kraftwerke ... ersatzlos stillzulegen.

Offensichtlich wird der Konjunktiv Imperfekt für potentiale Modalität, der Konjunktiv Plusquamperfekt dagegen für irreale Modalität verwendet. Der Zeitbezug spielt dabei nur eine untergeordnete Rolle. In diesem Falle hat die nächste Auflage des Duden (2005, §751) die empirische Erkenntnis aufgenommen, und die Beschreibung wurde verbessert.[11]

3.3 Zwischenbilanz

In den genannten Fällen hat sich die Grammatikschreibung jahrhundertelang im Kreise gedreht. Man fühlt sich an Brechts Galilei erinnert, der die Mitglieder seiner Fakultät auffordert, durchs Fernrohr zu gucken, um die tatsächlichen Verhältnisse zur Kenntnis zu nehmen, die jedoch abwinken, weil Erkenntnisse, die mit den Lehren des Aristoteles unvereinbar sind, von keinem wissenschaftlichen Belang sind. Datenbasiertes, induktives Vorgehen hat spekulative Grammatikerlehre ad absurdum geführt.[12]

Rückblickend kann man sagen, dass wesentliche Einsichten in das Funktionieren menschlicher Sprachen erst errungen worden sind, seit man unvoreingenommen hingeguckt hat, was für Äußerungen wirklich vorkommen und wie sie verwendet werden. Solche Einsichten betreffen nicht nur, wie man erwarten würde, die Semantik und Pragmatik, sondern sogar die grammatische Struktur. Und sie betreffen auch nicht nur die angemessene Analyse von bekannten Phänomenen, sondern z.T. sogar genuin neuartige Phänomene, von denen die Grammatiken und Lexika nichts wissen. Die kann man nur finden, wenn man unvoreingenommen an Texte herangeht. Die Konfrontation mit nicht selbst erzeugten Daten ist eine Quelle der Inspiration für den Linguisten, der nach Erkenntnissen strebt.

Selbstverständlich ist Korpuslinguistik nur eine neben anderen Methoden und nicht alleinseligmachend. Z.B. setzt die Beantwortung vieler linguistischer Fragen das Verfügen

---

[11] Leirbukt 2004 stellt sogar noch weitere Ausdehnung des Konjunktiv Plusquamperfekt fest.

[12] Sinclair (1991) schließt, dass Linguisten nun endlich dem Text mehr vertrauen können als ihren eigenen Intuitionen oder den auf ihnen basierenden allgemein akzeptierten grammatischen Beschreibungen.

über Minimalpaare und Paradigmen voraus, und die sind oft sehr viel leichter durch Informantenbefragung als durch Durchforstung von Korpora zu bekommen. Aber die darauf gegründete grundsätzliche Skepsis gegenüber Korpusarbeit verliert an Kraft, je umfangreicher die Korpora werden. Wenn eine Form in dem Korpus der Gesamtheit dessen, was ein Sprecher in seinem Leben gehört hat, nicht auftritt, dann bleibt ihm nichts anderes übrig, als sie in Analogie zu gehörten Formen zu bilden. Ebenso verhält sich der Linguist angesichts von endlichen Korpora; und je größer diese sind, desto wahrscheinlicher ist es, dass er im Einzelfall damit richtig liegt.

4 Dokumentation

4.1 Dokumentation und Beschreibung einer Sprache

Varietäten oder Genres einer Sprache werden schon seit langem dokumentiert, z.B. in umfassenden Editionen oder Anthologien. Die Idee, eine ganze Sprache zu dokumentieren, ist relativ neu. Sie ist eine Reaktion auf das erst in den achtziger Jahren des 20. Jahrhunderts erwachte Bewusstsein von der akuten Bedrohung der meisten Sprachen der Welt und von der Dringlichkeit, sie umfassend zu dokumentieren, falls sie dem Gedächtnis der Nachwelt nicht endgültig verloren gehen sollen.

Innerhalb der **umfassenden Darstellung einer Sprache** sind die Dokumentation und die Beschreibung aufeinander zugeordnet (s. Lehmann 2001, 2002). Die **Dokumentation** besteht i.w. aus einem Textkorpus, das gewissen Bedingungen genügt, auf die ich in § 4.3 komme. Die **Beschreibung** hat das Sprachsystem und die Situation der Sprache zum Gegenstand. Die Beschreibung verhält sich also zur Dokumentation so wie die *langue* zur *parole*. Eine Einheit des Sprachsystems steht zur einer Einheit einer Äußerung im Verhältnis des **Typs** (engl. *type*) zum **Vorkommen** (engl. *token*). Die Darstellung einer Sprache kann versuchen, dieses logische Verhältnis im formalen Aufbau von Beschreibung und Dokumentation nachzubilden. Die Beschreibung enthält dann Mengen von Objekten, deren jedes eine sprachliche Einheit repräsentiert und eine Identifikationsnummer hat. Das Korpus enthält nicht, wie natürlich gewachsene Korpora, die buchstäbliche Repräsentation sprachlicher Einheiten, sondern lediglich Ketten von deren Identifikationsnummern. Das Ganze lässt sich z.B. als objektorientierte relationale Datenbank implementieren. Formal betrachtet, liegt die Schnittstelle zwischen Beschreibung und Dokumentation dann in der Identifikation der Nummern zwischen den beiden Teilen. Falls das Korpus einmal jemand ansehen will, werden online zu den enthaltenen Vorkommen anhand der Identifikationsnummern die betreffenden Objekte des Systems herausgesucht und in der gewünschten Form, z.B. orthographisch oder phonetisch, dargestellt.

4.2 Sprachliche Variation zwischen Beschreibung und Dokumentation

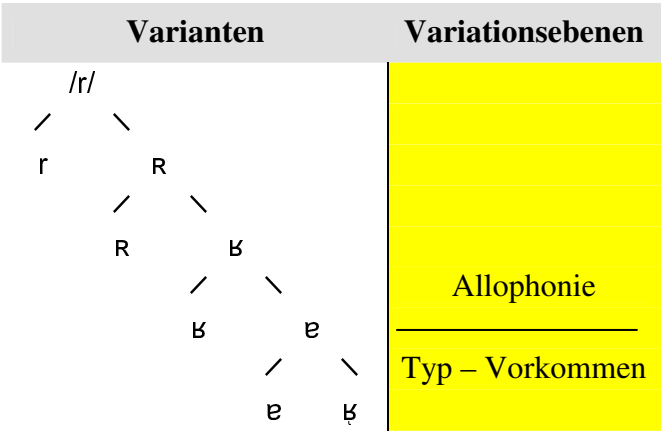
Eine solche Implementation stellt das Problem, wie **sprachliche Variation** zu erfassen ist, mit besonderer Klarheit. Denn die in den Rohdaten enthaltenen Vorkommen sprachlicher Einheiten sind selbstverständlich alle verschieden. Man kann aber nicht pro Textvorkommen einen Typ im System ansetzen, sonst wird das System infinit. Folglich muss man verschiedene Vorkommen auf einen Typ beziehen und nur diesen im Korpus repräsentieren,

muss also von gewissen Unterschieden zwischen Vorkommen im Korpus abstrahieren. Das bedeutet aber, dass man zwei Hauptebenen der sprachlichen Variation unterscheidet:

- Die obere Variationsebene ist Bestandteil des Sprachsystems oder der Sprachnorm. Die Varianten – z.B. Allophone und Allomorphe – sind Elemente der Sprachbeschreibung.
- Die untere Variationsebene gehört der Äußerung an. Die Varianten – z.B. idiolektale phonetische Feinheiten – werden übergangen und nirgends repräsentiert.

Das Schema Abb. 3 illustriert das Gemeinte am Beispiel des deutschen /r/. Eine enge phonetische Transkription als Repräsentation eines Textes des Korpus kann die Variation auf beliebigen Stufen der Feinheit wiedergeben. Tatsächlich wird aber eine Entscheidung gefällt, phonetische Variation unterhalb einer bestimmten Abstraktionsebene zu ignorieren. Man zieht also in die Variation eine Trennlinie ein.

Abb. 3. *Allophonie von /r/*



Die Variation oberhalb der Trennlinie ist Bestandteil des Systems (bzw., mit Coseriu, der Norm). Die Sprachbeschreibung enthält daher die beiden Allophone [ʁ] und [ɐ] (die z.B. an dritter Position in *Kurt* auftreten). Die Variation unterhalb der Trennlinie ist für das System irrelevant. Alle Varianten unterhalb der Trennlinie stehen in der Relation eines Vorkommens zu einem Typ des Systems, welcher ein Allophon der untersten Ebene ist. Die im Korpus tatsächlich vorgekommen [ʁ̞] werden als Vorkommen des Typs [ɐ] repräsentiert; dass sie tatsächlich [ʁ̞] waren, wird nie wieder jemand wissen.

Hieraus folgen zwei **Grundsätze für die Edition** eines Textkorpus:

- Vor Abschluss der Beschreibung der Sprache kann man nicht sicher sein, ob eine für irrelevant gehaltene Variation nicht doch eine Funktion hat oder in anderer Hinsicht theoretisch interessant ist. Deshalb ist es vorsichtiger, die Trennlinie möglichst niedrig anzusetzen, d.h. bei der Edition Variation nicht auszumerzen, sondern darzustellen. Es besteht keine Gefahr, den Wald vor lauter Bäumen nicht mehr zu sehen. Denn die Varianten, die man in der Korpusrepräsentation auseinandergehalten hat, sind in der Beschreibung des Systems auf eine Invariante bezogen. Jeder, der sich bloß für die Invariante interessiert, kann die Variation ausblenden. Aber wenn die Variation bereits durch indistinkte Repräsentation im Korpus ausgeblendet ist, ist sie der Forschung für immer verloren.
- Die Rohdaten, von denen das Korpus eine Repräsentation ist, sind aufzubewahren.

4.3 Repräsentativität und Qualität

Die beiden wichtigsten Anforderungen an ein Textkorpus, das zur Dokumentation einer Sprache dienen soll, sind Repräsentativität und Qualität (vgl. Lehmann 2001, § 5). Die **Repräsentativität** erfordert, dass die Variation auf folgenden Dimensionen im Korpus vertreten ist:

- Diastratik: Sprecher verschiedener Gruppen,
- Diatopik: Situationskontexte,
- Kommunikationsaufgaben und Textgenres,
- Themen,
- Medien.

Der letzte Punkt erfordert, dass mündliche Texte stärker vertreten sind als schriftliche, denn sie verwenden das fundamentale Kommunikationsmedium.[13]

Die genannten Dimensionen der Variation betreffen z.T. Außersprachliches, z.T. Pragmatisches und Semantisches. Der Systemlinguist will natürlich sicherstellen, dass auch **strukturelle Variation** herrscht, insbesondere, dass alles, was zum Sprachsystem gehört, im Korpus auch vertreten ist. Bei wohlbekannten Sprachen kann man auf Wege sinnen, dies gezielt sicherzustellen. Bei schlecht beschriebenen Sprachen ist es unmöglich, da man die Strukturen, die man repräsentiert sehen möchte, noch nicht kennt. Es besteht aber die Vermutung, dass wenn man für funktionale Variation auf all den genannten Dimensionen sorgt, sich die strukturelle Variation von selbst einstellen wird, nach dem Prinzip ‘form follows function’.

Die **Qualität** der Texte betrifft zunächst ihre Erzeugung, also die phonetische, grammatische, stilistische Kunst des Sprechers sowie den Inhalt der Nachricht. Dieses Erfordernis, das dem professionellen Linguisten seltsam vorkommen mag, wird den Philologen ebenso wie auch den Laiennutzer der Dokumentation wie eine Trivialität anmuten. Die Dokumentation wird für die Nachwelt gemacht; für minderwertige Produkte hat sie weder Interesse noch Ressourcen. Die Qualität betrifft aber ebenso die Arbeit des Kompilators des Korpus, also die Güte der Video- oder Audioaufnahme, der Transkription, der Annotation und der Metadaten. Qualität geht hier eindeutig vor Quantität.

5 Schlussfolgerungen

5.1 Normative Grammatik vs. tatsächlicher Sprachgebrauch

Seit der Antike war die Grammatik eine *ars*, also eine Kunst. Bis in die Neuzeit hinein gehörte Sprachwissenschaft zusammen mit den Philologien in die Faculty of Arts oder die Faculté des Lettres. Der normative Grammatiker beherrscht die Kunst des grammatischen Sprechens und Schreibens und vermittelt sie seinen Schülern so ähnlich, wie ein Klaviervirtuose diese Kunst seinen Schülern vermittelt. Diese Auffassung von Grammatik wird heute von keinem Sprachwissenschaftler mehr vertreten. Aber sowohl Grammatiken, die Richtschnuren für den korrekten Ausdruck liefern wollen, als auch Grammatiken, die

[13] Die Frage, wie die diachrone Variation einzubeziehen ist, kann hier nicht angemessen gewürdigt werden.

vorgeben, deskriptiv zu sein, werden de facto bis auf den heutigen Tag von Grammatikern im Vollgefühl persönlicher Sprachbeherrschung geschrieben.[14]

Eigentlich hat sich der Anspruch des Grammatikers in den letzten 200 Jahren sogar zugespitzt. Ein Musiker oder ein Maler sind zwar in der Lage, ihre Kunst den Gesellen zu vermitteln, erheben aber nicht den Anspruch, die gesamte Musik oder die gesamte Malerei einer Gesellschaft vollgültig zu repräsentieren. Der Grammatiker der Antike und des Mittelalters war hierin jedem anderen Künstler vergleichbar. Gerade diesen Alleinvertretungsanspruch haben Linguisten während der Herausbildung der strukturalen Sprachwissenschaft wenn auch nicht explizit, so doch immer selbstverständlicher erhoben. Dabei ist gerade in den Jahrzehnten der generativen Grammatik, als subtilste Konstruktionsunterschiede den Unterschied zwischen grammatisch und ungrammatisch ausmachen und somit den Ausschlag in theoretischen Entscheidungen geben sollten, sehr deutlich geworden, dass zwischen Grammatikern in bezug auf die Sprache, die sie zu beherrschen glauben, überhaupt keine Einigkeit herrscht. Es gibt folglich keine Basis für die Annahme, der Grammatiker sei allein imstande, den herrschenden Sprachgebrauch zu beurteilen. Es ist heute völlig klar, dass empirische Sprachbeschreibung die Auswertung von Korpora erfordert.

Andererseits stellt sich die Frage neu, wie eine normative Grammatik zu begründen ist. Wenn der einzelne Linguist hierfür keine relevante Instanz ist und die Gemeinde der Linguisten kein gemeinsames Urteil hat, auf welcher Grundlage kann man dann überhaupt Regeln für richtiges Reden und Schreiben formulieren? Wie es scheint, hat man in dieser Frage bescheiden auf die Stufe der Antike zurückzukehren, wo eine einzelne Person oder Gruppe durch ihren Sprachgebrauch einigen als Vorbild dient, anderen jedoch nicht. Irgendein allgemein-normativer Anspruch ergibt sich für diese Person oder Gruppe daraus nicht.

5.2 Sprachbeherrschung

Die generative Grammatik (Chomsky 1965:3f) hat den **idealen Sprecher-Hörer** postuliert, der seine Sprache vollkommen beherrscht und keinen Beeinträchtigungen der Performanz unterliegt. Dieser hat jedoch kein empirisches Korrelat. In der wissenschaftlichen Praxis sind die idealen Sprecher-Hörer überraschenderweise immer identisch mit den jeweiligen Linguisten. Der ideale Sprecher-Hörer entpuppt sich somit als methodischer Schachzug, um unter der Fahne empirischer Wissenschaft weiterhin normative Grammatik zu betreiben.

Tatsächlich gibt es in diesem Zusammenhang aber ein interessantes empirisches Feld, das viel zu wenig untersucht worden ist. Fasst man nämlich Kompetenz als **Sprachbeherrschung** auf, so wird es möglich, den Begriff zu operationalisieren. Man kann Kriterien für die Beherrschung einer Sprache definieren, ähnlich, wie das z.B. der Europarat (Council of Europe 2000, Goethe-Institut & Internationes 2001) für die Beherrschung von Fremdsprachen getan hat. Diese Aufgabe ist in jeder Hinsicht anspruchsvoll, denn die Kriterien sind sehr heterogen und unterliegen Einstellungen und Bewertungen der Sprachgemeinschaft, die ihrerseits wieder Gegenstand empirischer Untersuchungen sind. Zudem ist das ganze Thema ein heißes Eisen. Denn während das Faktum, dass Menschen F r e m d s p r a c h e n zu unterschiedlichen Graden beherrschen, jedem bekannt und auch in Ausbildungsinstitutionen als solches akzeptiert ist, wird über die u n t e r s c h i e d l i c h e B e h e r r s c h u n g

[14] S. Klein 2004 über Präskriptivität in modernen deskriptiven Grammatiken.

d e r M u t t e r s p r a c h e i.a. geschwiegen. Deutschlehrer und Germanistikprofessoren (um nur diese zu nennen) ahnen zwar dumpf, dass dieser Begriff ein empirisches Korrelat hat. Aber erstens sind die Kriterien der Beurteilung nie systematisiert worden. Zweitens ist der Vergleichsmaßstab i.a. die jeweilige Sprachnorm, und ein reflektierender Mensch hat Zweifel, ob der schiere Vergleich von jemandes Sprachbeherrschung mit der Sprachnorm zu einer bedeutsamen oder gar gerechten Beurteilung der ersteren ausreicht. Und drittens spielt bei Bewertungen immer auch die Ätiologie mit hinein: Wenn jemand eine Sprache weniger gut beherrscht, liegt das dann an minderer Begabung, an minderer bzw. anders orientierter Motivation, am sozialen Umfeld, an schlechter Ausbildung oder vielleicht noch anderen Faktoren? Wer hierüber empirisch forschen will, begibt sich auf vermintes Gelände.

Kein Systemlinguist würde seine introspektiven Urteile öffentlich damit rechtfertigen, dass er die Sprache zu einem Grade beherrscht, der die Beiziehung anderer Mitglieder der Sprachgemeinschaft erübrigt. De facto sind jedoch die Art und das Maß, in dem Linguisten ihre Sprache beherrschen, ihrerseits wieder empirischer Untersuchung zugänglich. Und wenn man die Frage unvoreingenommen angeht, ist das eine Untersuchung mit offenem Ausgang. Solange freilich diese Frage nicht empirisch untersucht ist, besteht für Linguisten kein Grund anzunehmen, ihre introspektiven Urteile könnten als Basis ihrer deskriptiven Aussagen dienen.[15]

5.3 Chancen und Gefahren

Daten erfüllen eine unentbehrliche Funktion in einer empirischen Wissenschaft. Aber sie sind kein Selbstzweck. Viele Untersuchungen, die unter der Flagge der Wissenschaft segeln, beschränken sich auf die Sammlung und Sortierung von Daten. Das ist kein spezifisch linguistisches Phänomen, sondern passiert in allen empirischen Wissenschaften. Viele Naturwissenschaften ertrinken in Daten. Satelliten nehmen Strahlung sämtlicher Wellenlängen in beliebig feiner Auflösung auf und funken die Werte Computern zu, die sie digitalisiert abspeichern. Es werden Löcher Hunderte von Metern tief in den Meeresboden gebohrt und das Profil der Sedimente über die Länge der Bohrsäule verzeichnet. Es gibt Karten, Diagramme, Kurven über alles, was sich aufzeichnen lässt. Ständig werden Technologien entwickelt und von Forschungsinstituten gekauft und eingesetzt, um noch mehr Daten aufzunehmen. Wenn irgendetwas technisch machbar ist, wird es gemacht. Es werden mehr Daten aufgenommen, als man je auszuwerten und für die Konstruktion von Theorien zu nutzen hoffen kann.

Viele philologische Dissertationen vom 18. bis ins 20. Jh. sind wenig mehr als Sammlungen von Belegstellen zu einem bestimmten Thema. Viele Arbeiten, die kommunikationsanalytisch sein sollen, gehen kaum über die Transkription von Gesprächen und deren Interpretation mithilfe gesunden Menschenverstandes hinaus. Wenn wir Korpuslinguistik betreiben und uns dafür einsetzen, dass in der Linguistik echte Daten die selbstersonnenen Beispiele ersetzen, so laufen wir Gefahr, als Faktenhuber missverstanden zu werden. Tatsächlich wäre für die Linguistik nichts gewonnen, wenn sich Korpuslinguistik auf Faktenhuberei reduzierte; und die rhetorische oder wissenschaftspolitische Abdrängung der Korpuslinguistik in diese Ecke wäre ebenso schädlich, weil sie einer falsch verstandenen theoretischen Linguistik Auftrieb

[15] In Kepser & Reis 2005 z.B. wird davon ausgegangen, es gäbe so etwas wie „introspective evidence“, aber diese Annahme wird an keiner Stelle plausibilisiert.

verschaffen würde. Es ist daher kein Zufall, dass die neue Zeitschrift der Disziplin sich *Corpus Linguistics and Linguistic Theory*[16] nennt und in ihrem Ankündigungstext verlangt, Beiträge müssten nicht nur "corpus-based", sondern auch "theoretically relevant" sein. Korpuslinguistik läuft nicht auf eine Senkung der theoretischen, sondern auf eine Hebung der methodischen Ansprüche hinaus.

Literatur

Chafe, Wallace 1992, "The importance of corpus linguistics to understanding the nature of language." Svartvik, Jan (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin & New York: Mouton de Gruyter (Trends in Linguistics); 79-97.

Chomsky, Noam 1965, *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press (Special Technical Report).

Corbett, Greville C. 2005, "Suppletion in personal pronouns: Theory vs. practice, and the place of reproducibility in typology." *Linguistic Typology* 9:1-23.

Council of Europe 2000, *A common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Du Bois, John W. 1987, "The discourse basis of ergativity." *Language* 63(4):805-855.

Duden 1984, *Grammatik*. Mannheim: Bibliographisches Institut.

Duden 2005, *Grammatik*. Mannheim: Bibliographisches Institut (7. Aufl.).

Givón, Talmy 1984, *Syntax. A functional-typological introduction.* Vol. I. Amsterdam & Philadelphia: J. Benjamins.

Goethe-Institut Inter-Nationes et al. (eds.) 2001, *Gemeinsamer Europäischer Referenzrahmen für Sprache Lernen, lehren, beurteilen*. München: Goethe-Institut (www.goethe.de/referenzrahmen).

Hempel, Carl Gustav 1966, *Philosophy of natural science*. Englewood Cliffs, N.J.: Prentice-Hall (Prentice-Hall foundations of philosophy series).

Kepser, Stefan & Reis, Marga 2005, "Evidence in linguistics". Kepser, Stefan & Reis, Marga (eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives*. Berlin: Mouton de Gruyter (Studies in Generative Grammar, 85); 1-6.

Klein, Wolf Peter 2004, "Deskriptive statt präskriptiver Sprachwissenschaft!? Über ein sprachtheoretisches Bekenntnis und seine analytische Präzisierung." *Zeitschrift für germanistische Linguistik* 32:376-405.

Lambrecht, Knud 1994, *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Lehmann, Christian 2001, "Language documentation: a program." Bisang, Walter (ed.), *Aspects of typology and universals*. Berlin: Akademie Verlag (Studia Typologica); 83-97.

Lehmann, Christian 2002, "Structure of a comprehensive presentation of a language. With particular reference to the interface between text, grammar and lexicon." Tsunoda, Tasaku (ed.), *Basic materials in minority languages 2002*. Osaka: Osaka Gakuin University (Endangered Languages of the Pacific Rim Publication Series); 5-33.

Lehmann, Christian 2004, "Data in linguistics." *The Linguistic Review* 21(3/4):275-310.

Leirbukt, Oddleif 1991, "*Nächstes Jahr wäre er 200 Jahre alt geworden*. Über den Konjunktiv Plusquamperfekt in hypothetischen Bedingungsgefügen mit Zukunftsbezug." *Zeitschrift für germanistische Linguistik* 19(2):158-193.

[16] Herausgegeben von Stefan Th. Gries und Anatol Stefanowitsch, publiziert von Mouton de Gruyter. Die Herausgeber postulieren ein "commitment to the systematic and exhaustive analysis of naturally occurring language".

Leirbukt, Oddleif 2004, "Über Konjunktiv Plusquamperfekt und *würde* + Infinitiv II als Ausdruck von Potentialität oder Irrealität in Konstruktionen mit Gegenwarts- oder Zukunftsbezug." Leirbukt, Oddleif (ed.), *Tempus/Temporalität und Modus/Modalität im Sprachenvergleich.* Tübingen: G. Narr (Eurogermanistik, 18); 205-230.

Oevermann, Ulrich 2002, „Klinische Soziologie auf der Basis der Methodologie der objektiven Hermeneutik – Manifest der objektiv hermeneutischen Sozialforschung“. Institut für objektive Sozial- und Kulturforschung (www.ihsk.de am 09.03.2006).

Poplack, Shana 2001, “Variability, frequency, and productivity in the irrealis domain of French.” Bybee, Joan & Hopper, Paul (eds.), *Frequency and the emergence of linguistic structure.* Amsterdam & Philadelphia: J. Benjamins (Typological Studies in Language, 45); 405-428.

Sinclair, J.M. 1991, *Corpus, concordance, collocation*. Oxford: Oxford University Press.