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 Korpusdaten[1] folgt.

In den darauf folgenden Abschnitten werden wir zwei für die Korpuslinguistik bedeutende sprachtheoretische Strömungen, die generative Grammatik und den Kontextualismus, vorstellen. Es geht dabei in erster Linie um den Platz von Korpusdaten in diesen Theorien.

[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.

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

1 Empirismus und Rationalismus

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

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

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

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

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

Betrachten wir ein Beispiel: In der Korpuslinguistik wurde in den 1990er Jahren der Begriff *Kollokation*[3] auf den Begriff der *Kookkurenz* (gemeinsames Vorkommen zweier linguistischer Einheiten, im Folgenden *Kovorkommen* genannt) zurückgeführt. Dem liegt die Einsicht zu Grunde, dass der Begriff der Kollokation nicht direkt auf Beobachtungen an Sprachdaten zurückzuführen ist. Es ist aber mittels Beobachtungen an Korpusdaten und statistischen Verfahren zu ermitteln, welche Paare

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

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

von Wörtern signifikant häufiger miteinander vorkommen, als dies auf Grund einer zufälligen Verteilung von Wörtern in Texten zu erwarten wäre. Mit Hilfe dieses nun auf Beobachtungen rückführbaren Begriffs des (signifikanten) Kovorkommens wurde der Begriff *Kollokation* neu definiert. Anders ausgedrückt: die Aussage, dass ein Wortpaar eine Kollokation ist, wird, da sie nicht direkt auf Erfahrung zurückzuführen ist, auf die Aussage gestützt, dass zwei Wörter signifikant häufig gemeinsam vorkommen, eine Aussage also, die direkt auf Erfahrung zurückführbar ist[4].

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

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

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

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

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

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

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

Aus der reinen Beobachtung und der Tatsache, dass Hauptsätze häufiger vorkommen als Nebensätze, könnte man nun schließen, dass das Deutsche tendenziell eine SVO-Sprache ist. In der generativen Grammatik wird statt dessen eine *Tiefenstruktur* angenommen, in der das Verb im Deutschen immer den Objekten folgt. In Hauptsätzen wird das finite Verb durch Transformationen oder vergleichbare Operationen an die zweite Position in der *Oberflächenstruktur* verschoben[8]. Es spricht einiges für eine solche Argumentation. Erstens kann auch in Hauptsätzen ein Teil des Verbalkomplexes hinter den Objekten stehen:

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

Zweitens wird die Partikel von Partikelverben dort quasi *zurückgelassen*:

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

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

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

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

[5] Vgl. Engfer 1996, S. 12.

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

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

[8] Im minimalistischen Programm (vgl. Chomsky (1995) und folgende) wird auf die Tiefenstruktur verzichtet, siehe z.B. Klenk (2003).

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

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

Die Deduktion lässt sich wie folgt charakterisieren:

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

Auch dies möchten wir an einem linguistischen Beispiel veranschaulichen: Aus der Beobachtung, dass *einige finite Verbformen Bestandteile von Hauptsätzen sind*, und der Beobachtung, dass *diese finiten Verbformen an zweiter Stelle im Satz stehen*, wird durch Induktion die Gesetzesaussage abgeleitet, dass *finite Verben in Hauptsätzen immer an zweiter Stelle stehen*. Diese kann an Beobachtungen überprüft und falsifiziert[9] werden. So trifft die Aussage z.B. für den Satz in Beispiel (5) nicht zu:

(5) Bleib wo du bist!

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

Anders herum kann aus dem unabhängig motivierten Prinzip der SOV- und SVO-Stellung von Konstituenten in Sätzen und der Feststellung, dass das Deutsche eine SOV-Sprache ist, deduktiv geschlossen werden, dass das finite Verb am Satzende steht. Die beobachtbare Tatsache, dass im Deutschen in Aussagesätzen das Verb an zweiter Stelle steht, wird mit der Regel dadurch in Einklang gebracht, dass eine Transformation angenommen wird, die das finite Verb aus der Endstellung in einer Tiefenstruktur an die zweite Position in der Oberflächenstruktur bewegt.

Im Rahmen rationalistisch orientierter sprachwissenschaftlicher Forschung kann ein Korpus zur Überprüfung und Korrektur theoretischer

[9] Mit *Falsifikation* wird das Verfahren bezeichnet, eine Gesetzesaussage durch mindestens ein Gegenbeispiel zu widerlegen bzw. zu verwerfen.

Aussagen verwendet werden. Wir werden dies *korpusgestützte* Linguistik nennen. Im Rahmen empiristisch orientierter sprachwissenschaftlicher Forschung ist das Korpus die primäre Quelle der Erkenntnis. Aus Beobachtungen an authentischen Sprachdaten werden Gesetzesaussagen abgeleitet, die durch weitere Beobachtungen bestätigt, modifiziert oder verworfen werden. Wir werden dies *korpusbasierte* Linguistik nennen[10].

2 Sprecherurteile statt Korpusdaten – Die Position der Generativen Grammatik

Alle sprachwissenschaftliche Forschung bezieht sich auf sprachliche Daten. Nur als eine Menge von gesprochenen oder geschriebenen Äußerungen kann sich das Sprachvermögen als kognitive Leistung von Menschen, oder das System einer natürlichen Sprache manifestieren. Schon Bloomfield stellte in den zwanziger Jahren des letzten Jahrhunderts in einem programmatischen Aufsatz fest, dass die Gesamtheit der Äußerungen, die in einer Sprachgemeinschaft gemacht werden können, die Sprache dieser Sprachgemeinschaft sei[11].

Bei dieser und bei ähnlichen Formulierungen zur Gegenstandsbestimmung der Sprachwissenschaft setzt nun die Kritik der generativen Grammatik[12] an, die seit den fünfziger Jahren das Forschungsprogramm der Sprachwissenschaft prägt. Die *Gesamtheit der Äußerungen* sei eine fiktive Größe, die im Fall einer lebenden, aktuell verwendeten Sprache durch keine Kollektion von Äußerungen auch nur annähernd repräsentiert werden könne. Eine Sprache durch Aufzählung aller Äußerungen erfassen zu wollen, sei nicht nur ein äußerst langweiliges, sondern auch ein müßiges Unterfangen. An dieser Stelle wird oft eine Analogie zum Schachspiel bemüht: Man lernt und versteht dieses Spiel nicht, wenn man die Zugfolgen möglichst vieler Partien betrachtet, sondern nur, indem man einige wenige Regeln lernt und diese anwendet. In ähnlicher Weise wird in der generativen Grammatik als eigentlicher Gegenstand der Forschung die kognitive Maschinerie (‚generative device') angese-

[10] Elena Tognini-Bonelli trifft eine ähnliche Unterscheidung und verwendet hierfür die Ausdrücke *corpus-based* linguistics und *corpus-driven* linguistics, vgl. Tognini-Bonelli 2001.

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

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

hen, die es Menschen ermöglicht, mit einem begrenzten Inventar von Regeln eine theoretisch unbegrenzte Menge von Äußerungen zu produzieren. Die Gesamtheit der bereits irgendwann getätigten Äußerungen sei für die Erkenntnis dieser kognitiven Maschinerie irrelevant.

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

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

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

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

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

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

So würde der tatsächlich belegte Satz:

(6) Anstelle des alten Magazins entstand vor einem Jahr ein fensterloser Trumm, *in dem* erst das Großkino „CinemaxX" einzog und nun auch das „Übermaxx" residiert ... (taz, 30. 4. 1999)

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

von den meisten deutschen Muttersprachlern, wenn er ihnen vorgelegt werden würde, als ungrammatisch empfunden – das Verb *einziehen* verlangt eine Präpositionalphrase im Akkusativ als Komplement, nicht eine im Dativ. Eine grammatische Beschreibung des Verbs *einziehen* würde aber, wenn Sie sich auf diesen Beleg stützte, eine Präpositionalphrase im Dativ als Komplement annehmen. Man könnte einwenden, dass die Beschreibung sprachlicher Phänomene sich nicht auf eine einzelne Beobachtung stützen sollte. Die Vorkommenshäufigkeit eines Phänomens spielt also eine wichtige Rolle.

Der Fehler im folgenden Beleg ist vermutlich kein Einzelfall:

(7) Allerdings haben die Bremer am 11. Mai noch ein Nachholheimspiel gegen Schalke 04, *daß* aus Sicherheitsgründen abgesagt wurde. (taz. 4. 5. 1999)

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

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

(8) *Peter wohnt.

(9) ?Peter wohnt mal wieder.

(10) Peter wohnt komfortabel.

(11) Peter wohnt in Berlin.

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

In späteren Arbeiten führt Chomsky eine weitere Unterscheidung ein, die zwischen E-Sprache und I-Sprache. Wir beziehen uns im Folgenden auf Chomskys Essay *Knowledge of Language. Its Nature, Origin, and Use*[14]. Chomsky charakterisiert sein Forschungsprogramm als Abstraktion weg vom konkreten sprachlichen Verhalten bzw. von dessen Produkten und hin zu den mentalen Zuständen, die dieses Verhalten bestimmen. Die Aufgabe der Sprachwissenschaft ist es, Antworten auf die folgenden Fragen zu finden:

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

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

Dem stellt Chomsky die *I-Sprache* (‚I-language', 22) gegenüber. Mit dem Ausdruck *internalisierte Sprache* bezeichnet Chomsky mentale Zustände der Sprecher, die eine Sprache beherrschen (22). Eine Grammatik ist eine Theorie über diese I-Sprache und damit über die mentalen Zustände der Sprecher. Grammatiken, verstanden als Theorien über die

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

I-Sprache, sollen so einfach wie möglich sein. Sie sind außerdem falsifizierbar wie jede andere wissenschaftliche Theorie. Dies sind die wissenschaftlichen Kriterien, nach denen Grammatiken bewertet werden können, wenn mehrere gleichermaßen die I-Sprache beschreiben. Die Konstruktion und Auswahl einer Theorie ist also keinesfalls willkürlich.

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

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

- Die wichtigste Quelle ist das Sprachgefühl bzw. Intuition (engl: ‚intuition') der Sprecher, die direkt oder indirekt über ihr Sprachwissen Auskunft geben[15]. Sprecher können direkt Auskunft geben, indem sie z.B. die Grammatikalität oder Akzeptabilität von Sätzen beurteilen, die ihnen vorgelegt werden, oder angeben, ob sie selber einen solchen Satz verwenden würden. Indirekte Auskunft kann dadurch eingeholt werden, dass Sprecher in Experimente einbezogen werden, in deren Verlauf ihnen bestimmte Äußerungen entlockt (engl. ‚elicit') werden[16].
- Darüber hinaus können die folgenden Quellen indirekt zu Erkenntnissen über das Sprachvermögen beitragen[17]:
  - Befunde über Sprachstörungen (Stottern, Aphasien u.s.w.);
  - Versprecher (*Sehr geehrte Hamen und Derren*)[18];
  - Neu geprägte Sprachen, z.B. die Kreolsprachen[19].

  Sprachstörungen sind ein indirekter Beleg für den modularen Aufbau des Sprachvermögens, denn bei den meisten Sprachstörungen sind nur einige Bereiche oder Aspekte des Sprechens gestört bzw.

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

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

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

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

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

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

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

- Korpora enthalten eine nicht unerhebliche Anzahl von Äußerungen, die von Sprechern, wenn sie diese Äußerungen zu beurteilen hätten, als nicht wohlgeformt eingestuft würden. Ursache für diese Einstufungen können banale Dinge wie Kongruenzfehler oder Wortauslassengen sein. Es kann sich aber auch um sehr subtile Phänomene handeln, deren (Nicht-)Wohlgeformtheit nicht einfach und einhellig festgestellt werden kann. Es sind diese subtilen (Pseudo-)Fehler, die die Interpretation von Korpusdaten besonders erschweren. Eine Grammatik einer Sprache, die sich ausschließlich, ohne ein weiteres Korrektiv, auf Korpusdaten dieser Sprache stützen würde, müsste solche Sätze wie in Beispiel (7) aufnehmen und grammatisch beschreiben[21].
- Selbst in den größten Korpora wird man eine Menge sprachlicher Phänomene, die für den Entwurf einer Grammatik der zu beschreibenden Sprache wichtig sind, nicht finden. Dies ist seit dem Aufkommen der generativen Grammatik eine Binsenweisheit. In jedem neuen Text wird man Sätze finden, die vorher noch nie geäußert bzw. aufgeschrieben wurden. Was für einzelne Sätze gilt, kann aber auch auf Konstruktionstypen zutreffen, und dieser Mangel ist für die Beschreibung von Sprachen oder gar für die Theoriebildung viel

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

[21] Ein weiteres, berühmtes Beispiel ist der Satz *Ich habe fertig*, geäußert vom italienischen Trainer Giovanni Trappatoni auf einer Pressekonferenz. Die Ursachen für diesen Fehler liegt in der mangelnden Beherrschung der deutschen Sprache. Der Satz hat aber durch häufige Pressezitate mittlerweile den Status eines geflügelten Wortes. Man wird ihm in Zeitungstexten dieser Zeit sicher häufig begegnen. Aber will man diesen Satz wirklich als *wohlgeformt* akzeptieren und beschreiben?

gravierender. Wenn in einem Korpus der deutschen Sprache keine Imperativform (wie z.B. *Gib!*) oder keine Mittelkonstruktion (wie z.B. *Dieses Auto fährt sich gut*) auftauchte, dann könnten diese Konstruktionstypen auch nicht in einer rein empirischen, korpusbasierten Sprachbeschreibung auftauchen. Sprecher des Deutschen würden diese Konstruktionen, wenn man sie ihnen vorlegte, aber sicher als wohlgeformt einstufen und sie bei gegebenem Anlass auch selber verwenden. Ihr Fehlen im Korpus ist ein reines Zufallsprodukt.

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

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

(12) John is smoking a lot anymore.

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

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

[22] Das Beispiel stammt aus McEnery und Wilson 1996, S. 11.

[23] Cf. Labov 1975, S. 34f.

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

[25] Einige wichtige Studien werden in Labov 1975 diskutiert.

manten gelten können, sind auch sie nicht vor Fehlurteilen sicher, wie das obige Beispiel zeigt[26].

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

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

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

Diese kritische Bewertung introspektiver Daten als Quelle linguistischer Erkenntnis soll nicht davon ablenken, dass auch Korpusdaten

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

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

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

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

[30] Wir empfehlen dem interessierten Leser drei neuere und interessante Arbeiten zu diesem Thema: Blume (2004), Featherston (2002) und Sorace und Keller (2005).

problematisch sein können. Die Kritik an dem Wert von Korpusdaten sei hier noch einmal in vier Punkten zusammengefasst:

1. Der Status eines beliebig großen Korpus zu der Sprache, die es repräsentieren soll, ist unklar, da die repräsentierte Sprache aus einer potenziell unendlichen Menge von Sätzen besteht (Problem der Repräsentativität);
2. Ein Korpus enthält eine große Zahl von Phänomenen, die für die Beschreibung der Sprache, die es repräsentiert, irrelevant sind (Problem der Relevanz der Daten);
3. Viele Konstruktionen, die im Beschreibungsbereich einer Grammatik liegen, da sie wohlgeformt sind, sind in Korpora dieser Sprache nicht vorhanden (Problem unvollständiger Datenabdeckung);
4. Viele Äußerungen, die dann auch Bestandteile von Korpora sein können, sind nicht wohlgeformt. Aus ihnen können und sollten keine Schlüsse auf das sprachliche Wissen der Sprecher gezogen werden (Problem der Verlässlichkeit der Daten).

Man sollte als Sprachwissenschaftler, der mit Korpora arbeitet, diese Kritik an Korpusdaten ernst nehmen und dieser Kritik mit guten Argumenten begegnen können. Hierzu gehören Antworten auf die Fragen, wie man mit der Existenz nicht-wohlgeformter Äußerungen und mit dem Fehlen wohlgeformter Äußerungen umgeht. Zweifelhafte Schlüsse können zum Beispiel durch andere Daten, wie Sprecherbefragungen, gestützt werden. Hierzu gehört auch die Antwort auf das Problem, das mit der Generativität von Sprachen zu tun hat: Wie bestimmt man das Verhältnis einer endlichen Menge von Beobachtungsdaten zu einer potenziell unendlichen Menge von Äußerungen? Zu diesem Problem findet man in der modernen Statistik einige überzeugende Antworten.

Wir werden in den folgenden Kapiteln zeigen, dass, bei aller Kritik, auch Sprachwissenschaftler, die in diesem theoretischen Rahmen arbeiten, auf Korpusdaten zurückgreifen. Dies ist in den letzten Jahren sogar vermehrt der Fall. Korpuslinguistik, und damit auch die methodische Diskussion, gewinnt auch hier an Relevanz.

Der nächste Abschnitt ist einer linguistischen Schule gewidmet, für die die Arbeit mit Korpusdaten notwendiger Bestandteil linguistischer Erkenntnis ist. Für diese Schule, die in Deutschland *Kontextualismus* genannt wird[31], geht alle linguistische Erkenntnis vom Sprachgebrauch aus.

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

3 Linguistische Erkenntnis geht vom Sprachgebrauch aus – Die Position des Kontextualismus

Einige prominente Korpuslinguisten, zum Beispiel John Sinclair – ehemaliger Chefredakteur des *Collins Cobuild English Dictionary* – entstammen der Schule des Kontextualismus. Für diese Sprachwissenschaftler ist die Arbeit mit sehr großen Textkorpora der Vollzug des Forschungsprogamms, das vor allem von John Rupert Firth entworfen wurde[32].

Das Forschungsziel des Kontextualismus ist es, sprachliche Äußerungen und deren verschiedene linguistische Aspekte als Funktionen des sprachlichen und nicht-sprachlichen Kontextes zu erklären, in dem diese Äußerungen stehen.

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

Auch die Kontextualisten möchten letztendlich zu Aussagen über das Sprachsystem gelangen. Soweit scheinen der Kontextualismus und die generative Grammatik übereinzustimmen. Ein wesentlicher Unterschied liegt allerdings im Verständnis dessen, was als *Sprachsystem* bezeichnet wird. Für die generative Grammatik ist dies eine kognitive Struktur, das sprachliche Wissen der Sprecher. Für den Kontextualismus sind dies die regelhaften Beziehungen zwischen der Form, dem Inhalt und dem Kontext sprachlicher Äußerungen. Diese Beziehungen können nur aus konkreten Sprachhandlungen abstrahiert werden. Am Beginn dieser Abstraktion muss also die Erfassung, Analyse und Systematisierung der konkreten Sprachhandlungen stehen. Für Firth liegt die Bedeutung linguistischer Einheiten in ihrer Funktion für den *Kontext*, in den die Äußerungen eingebettet sind.

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

Definition 3 (Kontext). *Der Kontext einer Äußerung ist die Summe der unmittelbaren Rahmenbedingungen einer Sprachhandlung als das Bezugssystem, innerhalb dessen einer Äußerung eine Funktion zukommt. Dabei bildet der* kulturelle Kontext *das Bezugssystem für eine Sprache und steuert die Art und Weise, wie Sprecher sprachliche Handlungen wahrnehmen. Der* situative Kontext *determiniert die Funktion einer konkreten sprachlichen Handlung. Zum situativen Kontext gehören Ort und Zeit, die Beteiligten, etc.*[33]

Das Konzept des Kontexts als Rahmen und Bedingung menschlichen Handelns hat Firth von dem Anthropologen Bronislaw Malinowski übernommen. Er hat dieses System auf linguistische Untersuchungen hin erweitert, indem er dem – im Wesentlichen nicht-sprachlichen – Kontext das Konzept des Kotextes an die Seite stellte[34].

Definition 4 (Kotext). *Der Kotext einer linguistischen Einheit ist die Menge der linguistischen Einheiten, die im gleichen Text verwendet wurden. Diese linguistischen Einheiten determinieren die Funktion und die Bedeutung der untersuchten Einheit.*

Ko- und Kontext spielen für die Untersuchung sprachlicher Handlungen eine zentrale Rolle. Sie haben den (deutschen) Namen dieser linguistischen Schule geprägt.

Firth hat den Kotext von Wörtern und Sätzen auf den vier Ebenen der Phonetik und Phonologie, der Morphologie, der Syntax und der Lexik untersucht. Die Untersuchungsbasis bildeten einzelne, situationsgebundene Texte. Heutzutage findet man natürlich eine große Zahl von Sprachhandlungen in Korpora dokumentiert, dies war aber zu Firths Zeiten noch nicht der Fall[35].

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

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

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

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

Definition 5 (Kollokation). *Innerhalb des Kontextualismus wird unter Kollokation das faktische Miteinandervorkommen zweier oder mehrerer beliebiger Wörter oder lexikalischer Einheiten verstanden. Damit ist keine normative Bewertung hinsichtlich der Korrektheit oder Grammatikalität dieser Wortverbindung verbunden. Der Begriff wird vom späten Firth und einigen seiner Anhänger auf die* Habitualität *des Kovorkommens eingeschränkt. Darunter wird vor allem verstanden, dass die Wortverbindung in den beobachteten Texten wiederholt auftreten muss*[36].

Die Analyse von Kotext und Kontext linguistischer Einheiten ist für Firth und seine Anhänger der Schlüssel zur Bedeutung dieser linguistischen Einheiten. Bedeutung wird also nicht, wie in vielen anderen Theorien, als eine mentale Disposition von Sprechern oder als eine Struktur, die unabhängig vom Gebrauch existiert, aufgefasst. Damit ist der Kontextualismus eine Gebrauchstheorie der Bedeutung, im Sinne von Wittgensteins berühmter Formel: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"[37]. Firth formuliert dies ganz ko(n)texutalistisch: „You shall know a word by the company it keeps"[38].

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

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

Definition 6 (Kolligation). *Als Kolligationen werden Paare sprachlicher Einheiten bezeichnet, deren Zusammenhang durch die Bezeichnung ihrer syntaktischen Kategorien und der Beziehungen zwischen diesen Kategorien weiter qualifiziert ist*[39].

Nach dieser Definition ist das Beispiel (13) allein auf Grund des häufigen Kovorkommens als Kollokation auffassen, nicht aber als Kolligation. Im

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

[37] Vgl. Wittgenstein (1967), S. 43.

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

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

Gegensatz dazu ist Beispiel (14) eine Kolligation, da zwischen den beiden Elementen die grammatische Beziehung von Prädikat und Objekt besteht.

(13) und er

(14) Antrag stellen

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

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

- Der Kontextualismus ist eine sprachwissenschaftliche Richtung, die linguistische Erkenntnis einzig und allein auf die Analyse des Sprachgebrauchs stützt. Die materielle Basis der linguistischen Untersuchungen, Texte und heutzutage Korpora, werden exhaustiv untersucht. Es werden von vornherein keine Daten ausgeschlossen (etwa, weil sie nicht wohlgeformt wären).
- Der bedeutendste Beitrag des Kontextualismus für die moderne Korpuslinguistik liegt in der Analyse von Wortverbindungen. Dabei dominiert der syntagmatische Aspekt, das gemeinsame Vorkommen der Wörter in einer größeren linguistischen Einheit, bei weitem den paradigmatischen Aspekt, der im Kontextualismus auch eine Rolle spielt[40]. Wortverbindungen können, je nach dem Status der Interpretation, als *Kollokationen* oder als *Kolligationen* bezeichnet werden.
- Die Analyse von Korpora im Geiste des Kontextualismus hat vor allem im Bereich der Lexikographie und Lexikologie, in der Übersetzungswissenschaft, für den Sprachunterricht und als Basis von sprachkritischen Untersuchungen bedeutende Leistungen ermöglicht.

Generative Grammatik und Kontextualismus unterscheiden sich, wie wir gesehen haben, hinsichtlich der Auffassung ihres Untersuchungsgegenstandes, hinsichtlich dessen, was als sprachliche Daten von Rele-

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

vanz für die Bildung abstrakter und generalisierter Aussagen über den Gegenstand ist, und dementsprechend auch hinsichtlich der Verwendung von linguistischen Korpora. Ein Austausch zwischen diesen beiden großen Strömungen in der modernen Linguistik fand bisher kaum statt. Erkenntnisse etwa über das kollokative oder funktionale Spektrum lexikalischer Einheiten werden von generativen Grammatikern als trivial und für eine ernsthafte Sprachtheorie irrelevant abgetan. Auf der anderen Seite werden von den Kontextualisten theoretische Aussagen der generativen Grammatiker als unbegründet, da empirisch nicht fundiert oder gar von jeglicher empirischer Basis isoliert und damit empirisch nicht falsifizierbar abgetan.

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

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

4 Korpusbasierte Ansätze

Die in den letzten Abschnitten beschriebenen methodischen Begriffspaare *Empirismus / Rationalismus* und *Deduktion / Induktion* gliedern die folgende Übersicht, siehe auch Tabelle 2. Wir unterscheiden drei Ansätze in der Korpusanalyse: den korpusbasierten, quantitativen Ansatz, den korpusbasierten, quantitativ-qualitativen Ansatz und den korpusgestützten, qualitativen Ansatz.

4.1 Der korpusbasierte, quantitative Ansatz

Bei diesem Verfahren werden auf der Grundlage von rohen, also nicht linguistisch annotierten, Korpora quantitative Daten extrahiert.

	korpusbasiert quantitativ	korpusbasiert, quantitativ und qualitativ	korpusgestützt
Ansätze	Latent-semantische Analyse N-Gramm Analyse	Kollokationsanalyse Semantische Prosodie	Wortstellungsphänomene
Theoretischer Rahmen	Keiner	Kontextualismus (Firth)	Strukturalismus (Saussure), Generative Grammatik (Chomsky)
Erkenntnistheoretischer Ansatz	extrem empiristisch	gemäßigt empiristisch	rationalistisch
Personen	Landauer – Jelinek	Sinclair, Teubert, Heringer	Fillmore, Arts, Oostdijk, Reis, Meurers
Eingabe	Korpus in Rohform	Korpus in Rohform	Linguistisch annotiertes Korpus oder Belegsammlung
Ausgabe	Text-Term-Matrizen N-Gramme mit Frequenzen	Kollokator-Kollokant-Paare mit Kennziffern	Belegsätze
Interpretation	Keine	Ja, von den Belegen ausgehend	Ja, von den theoretischen Aussagen ausgehend
Linguistische Domäne	statistische Sprachmodelle	Semantik	Syntax
Anwendungsgebiet	Informationserschließung, Verarbeitung gesprochener Sprache	Lexikographie, Sprachunterricht, Übersetzungswissenschaft	Lexikographie, theoretische Linguistik

Tabelle 1: Ansätze in der Korpuslinguistik

Diese quantitativen Daten können qualitativ interpretiert werden, dies ist aber für den erfolgreichen Einsatz dieser Verfahren nicht notwendig. Typische Kennziffern einer quantitativen Korpusanalyse sind:

- die absolute Häufigkeit, mit der eine Zeichenkette[41] in einem Text / Korpus vorkommt;
- die relative Häufigkeit, mit der eine Zeichenkette in einem Text / Korpus vorkommt;
- der Rangplatz, den eine Zeichenkette auf Grund ihrer Häufigkeit einnimmt (z.B. *Zeit* ist das zehnthäufigste Wort = das Wort *Zeit* hat den Rangplatz 10);
- die Distribution eines Wortes, gemessen als die Häufigkeit des Vorkommens dividiert durch die Zahl der Texte, in denen das Wort vorkommt;
- Häufigkeiten von Sequenzen anderer linguistischer Einheiten, beschrieben als *n-Gramme*, in Texten;
- semantische Ähnlichkeit von Wörtern, gemessen an der Häufigkeit ihres Kovorkommens oder gemeinsamen Vorkommens mit weiteren Wörter (s. Exkurs).

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

Exkurs: Quantitatve Verfahren im Information Retrieval

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

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

n-Gramme Vorkommenshäufigkeiten von *n-Grammen* linguistischer Einheiten können dazu verwendet werden, formähnliche Wörter in Anfrage und Text auf einander abzubilden. Es kann sich dabei um Folgen von 1, 2, 3 ... n Phonemen, Graphemen etc. handeln. Nehmen wir an, dass in einem Text *Operationen am offenen Herzen* vorkommt. In der Suchanfrage wird der Term *Herzoperation* verwendet. Bei einfachem Abgleich der Wörter würde das Dokument, das doch immerhin relevant erscheint, nicht gefunden. Beide Zeichenketten haben aber acht Trigramme gemeinsam: *_He, Her, erz, per, era, ati, tio, ion*, also ca. 90 Prozent der Trigramme der kürzeren Zeichenkette. Das n-Gramm-Verfahren ist eine Möglichkeit, der Schreibvarianten bei vielen Wörtern und Termen Herr zu werden. N-Gramm-Modelle werden ausführlich in Manning und Schütze (1999), Kap. 6 behandelt.

Latent-semantische Analyse Um semantisch ähnliche Wörter in Anfrage und Dokumenten aufeinander abzubilden, wird aus dem Vorkommen von Termen in Dokumenten deren Ähnlichkeit bestimmt. Ist ein Term in der Anfrage einem Term in einem Dokument semantisch ähnlich, dann steigert dies die Relevanz dieses Dokumentes für die Anfrage. Um die semantische Ähnlichkeit zwischen Termen zu bestimmen, werden Term-Dokument-Matrizen erzeugt. Diese Matrizen werden in wohldefinierter Weise manipuliert. Aus den manipulierten Matrizen lassen sich sowohl die Bedeutungsähnlichkeit von Termen als auch die Bedeutungsähnlichkeit von Texten bestimmen. Das Verfahren wird *latent-semantische Analyse* genannt.

Eine Matrize ist eine Tabelle mit Spalten und Zeilen. Jedes Wort nimmt eine Spalte ein und jeder Text eine Zeile. Die folgenden, sehr kurzen Texte:

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

(16) Terroristen verüben Anschlag (Text 2)

können wie folgt repräsentiert werden:
Miliz(1,0)
verhaftet(1,0)
Terroristen(1,1)
nach(1,0)
Anschlag(1,1)
verüben(0,1)
Hat man viele Dokumente und damit viele verschiedene Wortformen, dann entsteht eine sehr große Matrize (mit m Zeilen für m Wortformen und n Spalten für n Dokumente. Die Matrix enthält viele Leerstellen, da die meisten Wörter in den meisten Texten nicht vorkommen.

Die Singulärwertzerlegung als mathematische Operation über Matrizen bietet die Möglichkeit, solche großen Matrizen auf einige Hundert Dimensionen (= Zeilen und Spalten) zu verkleinern, bei optimaler Erhaltung der in ihnen kodierten Informationen. Zum mathematischen Hintergrund vgl. Berry, Drmac und Jessup (1999), wir können im Rahmen dieser Einführung nicht weiter darauf eingehen.

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

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

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

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

Zur Einführung in die latent-semantische Analyse empfiehlt sich die Lektüre von Landauer und Dumais (1997) (theoretischer Hintergrund) und Landauer, Foltz und Laham (1998) (Anwendungen).

4.2 Der korpusbasierte, quantitativ-qualitative Ansatz

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

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

Der größte Nutzen dieser Art von Korpuslinguistik konnte bisher in der Lexikographie, in der Übersetzungswissenschaft und für den Sprachunterricht erzielt werden[42]. Auch für sprachkritische Untersuchungen erwies sich der Ansatz als fruchtbar.

4.3 Der korpusgestützte Ansatz

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

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

Der größte Nutzen dieser Art von Korpuslinguistik konnte bisher dort erzielt werden, wo syntaktische Theorien auf Grund eines relativ klaren Befundes an authentischer Sprachverwendung falsifiziert wurden[43]. Wir werden dies an einigen Beispielen in den nächsten Kapiteln zeigen.

[42] Vgl. z.B. die Fallstudien in Tognini-Bonelli (2001).
[43] Vgl. z.B. Sampson (1996).

5 Weiterführende Literatur

Der theoretische Hintergrund der modernen Korpuslinguistik wird leider nur sehr selten thematisiert. Ein paar Seiten hierzu finden sich bei McEnery und Wilson (1996), Kapitel 1. Wolf Paprotté geht in zwei Aufsätzen etwas genauer auf diese Fragen ein (Paprotté, 1992, 1994). Ein lebendiges Bild der linguistischen Szene und ihrer Kämpfe vermittelt das Buch „The linguistics wars" (Harris, 1995). Dieses Buch eignet sich aber eher als Bettleküre denn als Referenzwerk, da es nur unzureichend erschlossen ist. Über die Positionen des Kontextualismus gibt Tognini-Bonelli (2001) Auskunft. Von diesem Buch haben wir die Unterscheidung in korpusbasierte und korpusgestützte Untersuchungen übernommen. Die Autorin argumentiert allerdings ganz aus der Sicht des Kontextualismus und ist insofern anderen Ansätzen gegenüber nicht immer ganz fair. In einem von Svartvik heraugegebenen Band eines hochkarätigen Symposiums (Svartvik, 1992) werden methodische Fragen reflektiert. Einige dort versammelte Aufsätze aus diesem Band sind aus dieser Perspektive besonders ergiebig (vor allem Fillmore, 1992; Chafe, 1992; Halliday, 1992; Leech, 1992).

6 Aufgaben

Hinweis: Lösungsansätze zu den Aufgaben finden Sie auf unserer Website (`http://www.lemnitzer.de/lothar/KoLi`).

1. Welche der folgenden Aussagen sind empirisch begründet und welche rationalistisch:
   a) Der Satz *Wo sollen wir treffen?* ist ungrammatisch.
   b) Sätze wie *Wo sollen wir treffen?* sind typische Fehler englischer Lerner des Deutschen, die das Verb *treffen* verwenden wollen.
   c) Der Kopf einer Nominalphrase ist das Nomen.
   d) Instruktive Texte (z.B. Kochrezepte) enthalten überdurchschnittlich viele Befehlsformen.
   e) Es gibt 15 Dialekte im Deutschen.
   f) In der Tiefenstruktur des deutschen Satzes steht das finite Verb am Satzende. Bei einigen Satztypen wird es bei der Realisierung der Oberflächenstruktur an Zweitposition verschoben.
   g) Kollokationen sind Paare von Wörtern, die überdurchschnittlich häufig miteinander vorkommen.
   h) *Hartes Leben* ist eine Kollokation.

   Für die Beantwortung welcher Frage benötigen Sie ein Sprachdatenkorpus?

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