Erschienen in: Kallmeyer, Werner/Zifonun, Gisela (Hrsg.): Sprachkorpora. Datenmengen und Erkenntnisfortschritt. – Berlin, New York: de Gruyter, 2007. S. 49-69. (Institut für Deutsche Sprache. Jahrbuch 2006), https://doi.org/10.1515/9783110439083-005

SAM FEATHERSTON

Experimentell erhobene Grammatikalitätsurteile und ihre Bedeutung für die Syntaxtheorie*

Abstract

In diesem Beitrag versuchen wir darzulegen, unter welchen Umständen introspektive Urteile objektive, quantifizierbare, und empirisch adäquate linguistische Daten sein können. Dazu skizzieren wir, wie die Resultate unserer experimentell erhobenen, relativen Urteilsstudien aussehen, und argumentieren, dass sie eine unverzichtbare Evidenzquelle für die Syntax bilden, weil sie Einsichten in die Architektur der Grammatik erlauben, die mit anderen Mitteln nicht möglich sind.

Die Jahrestagung 2006 des Instituts für Deutsche Sprache fand zum Thema ‚Sprachkorpora – Datenmengen und Erkenntnisfortschritt' statt. Das Projekt A3 ‚Suboptimale syntaktische Strukturen' im SFB 441 ‚Linguistische Datenstrukturen' in Tübingen betreibt Forschung zu Datentypen und Syntaxtheorie, wobei der Hauptakzent auf introspektiven Urteilen liegt. Durch die Erhebung der Urteile von Muttersprachlern des Deutschen zu grammatischen, marginalen sowie ungrammatischen Sätzen gewinnt man wichtige Indizien zu zentralen linguistischen Fragestellungen, wie zum Beispiel was es psycholinguistisch bedeutet, wenn eine Struktur als ‚gut' oder ‚schlecht' bewertet wird, und welche linguistisch relevanten und irrelevanten Faktoren bei dieser Differenzierung eine Rolle spielen. Für die Auswertung dieser Studien werden auch Frequenzdaten aus Korpora herangezogen, und deren Ähnlichkeiten mit und Unterschiede zu Urteilsdaten erforscht und ergründet. Gerade unser Projektschwerpunkt im Bereich der Datentypen erlaubt es, den Stellenwert von Frequenzdaten mit einer gewissen ‚Außensicht' zu kommentieren. Die Aufgabe unseres Beitrags in dem vorwiegend korpusorientierten Kontext dieses Bandes ist es deshalb, den Evidenzwert, sowie die Vor- und Nachteile von Frequenzdaten relativ zu anderen Datentypen zu kommentieren, um erstens die zum Teil komplementären Stärken und Schwächen von Frequenzdaten und Urteilen zu fokussieren, und zweitens die Einsichten darzustellen, die ausschließlich mithilfe von Urteilsdaten zu gewinnen sind.

* Diese Arbeit entstand im Rahmen des Projekts A3 ‚Suboptimale syntaktische Strukturen' im SFB 441 ‚Linguistische Datenstrukturen' in Tübingen. Die Unterstützung der Deutschen Forschungsgemeinschaft wird dankend anerkannt. Vielen Dank an Wolfgang Sternefeld und Tanja Kiziak aus dem Projekt und den Entwicklern von *WebExp*.

Dieses Papier hat deshalb zum Ziel, zwei Thesen aufzustellen und deren Wert zu belegen. Die Thesen lauten:

1. Introspektive Urteile können objektive, quantifizierbare Daten sein.
2. Introspektive Urteile sind auch für den Syntaktiker notwendig, um ein volles Verständnis der Syntax zu erlangen.

Im Folgenden werden wir zuerst die Gründe für die Zweifel anführen, die Sprachwissenschaftler zunehmend dazu bewegen, Misstrauen gegenüber Urteilen als Datentyp zu hegen, und auf andere Datentypen, wie zum Beispiel Frequenzdaten, auszuweichen. Wir werden die Motiviertheit dieser Zweifel nicht bestreiten, sondern darlegen, wie diese Schwächen des Datentyps gemildert und vermieden werden können. Im Wesentlichen geht es darum, statt Urteile als Einzelperson selbst abzugeben, die Daten von Informantengruppen unter Einhaltung verschiedener Gebote der experimentellen Kontrolle zu benutzen. Wir werden argumentieren, dass methodologisch einwandfrei erhobene introspektive Daten durchaus als empirisch adäquat gelten können und darüber hinaus spezifische Vorteile in ihrem Evidenzwert haben.

In einem zweiten Schritt werden wir auf die Befunde von Studien mit dieser Methode eingehen. Tatsächlich liefern diese *relativen Urteile* ein anderes Bild von der Funktionsweise der Syntax als herkömmlich angenommen wird. Statt eines kategorischen Modells der Grammatikalität unterstützen diese Daten vielmehr ein Modell der gradierten Wohlgeformtheit. Insofern stimmen die Erkenntnisse aus diesen experimentell erhobenen Urteilen mit dem Muster überein, das man von Frequenzdaten kennt. Es gibt aber auch Parameter, bei denen sich die zwei Datentypen unterscheiden. Insbesondere liefern Häufigkeiten und Urteile unterschiedliche Verteilungsmuster. Häufigkeitsdaten weisen ein gradiertes Bild der Wohlgeformtheit auf, weil sie eine Vielzahl von binären Entscheidungen summieren; jedes einzelne relative, in numerischer Form abgegebene Urteil dagegen beinhaltet Gradiertheit bereits in sich selbst. Aus dieser Tatsache kann man Schlüsse über die Architektur der Teile des menschlichen syntaxverarbeitenden Prozesses ziehen, die auch für das Zusammenwirken der verschiedenen Funktionen in der Grammatik wichtig sind. Zur vollen Einsicht in die Architektur der Grammatik kommt man nur, indem man sowohl Frequenzdaten als auch Urteilsdaten berücksichtigt. Daher ist unsere Schlussfolgerung, dass sowohl Frequenzen wie auch introspektive Urteile für die Theoriebildung in der Syntax nötig sind, jedoch erlauben Urteile einen leichteren Zugang zu den einzelnen syntaktischen Beschränkungen als Frequenzdaten.

Introspektive Urteile als Datentyp

In diesem Teil werden wir argumentieren, dass introspektive Urteile unter bestimmten Umständen als ein verlässlicher, objektiver und quantifizierbarer Datentyp gelten können.

Introspektive Urteile: die Nachteile

Es ist immer wieder in der Literatur kritisiert worden, dass introspektive Urteile als Datentyp deutliche Nachteile haben; zum Beispiel von Labov (1975) und Sampson (2001), siehe Schütze (1996) für eine ausführliche Diskussion. Es wird Urteilen erstens vorgeworfen, dass sie ungenau sind. Das ist natürlich für das einzelne Urteil richtig: stellt man verschiedenen Informanten die gleiche Frage, dann geben sie öfter verschiedene Antworten. Man kann auch derselben Person die gleiche Frage zu verschiedenen Zeitpunkten stellen und dann nicht selten unterschiedliche Ergebnisse erhalten. Zum Teil als Reaktion auf dieses Phänomen haben Syntaktiker daher oft zum einen mit der Idee der individuellen Idiogrammatik gearbeitet und zum anderen dazu tendiert, Nicht-Linguisten die Fähigkeit abzusprechen, die Feinheiten der Urteilsabgabe ausreichend zu beherrschen.

Ein weiterer Vorwurf ist die mangelnde Quantifizierbarkeit von Urteilen. Dies hat zur Folge gehabt, dass manche Autoren nur binär zwischen guten und schlechten Beispielen unterscheiden, während andere ‚gut', ‚marginal' und ‚schlecht' differenzieren, in extremen Fällen gibt es bis zu sieben abgestufte Grade der Wohlgeformtheit (zum Beispiel hat Müller 1995 fünf: Ø [keine Angabe = ‚vollgrammatisch'], ?, ??, *?, und *. Lakoff 1973 sechs: Ø, ?, ??, ?*, * und **. Wurmbrand 2001 sogar sieben: Ø, #, %, ?, ?, ??, *). Sprecher haben ein intuitives Gefühl dafür, was absolut grammatisch oder ungrammatisch ist, aber man kann bezweifeln, ob das ??-Urteil von Lakoff dem ??-Urteil von Wurmbrand entspricht, wenn man bedenkt, dass ?? von Lakoff die drittbeste von sechs Stufen ist und ?? von Wurmbrand die zweitschlechteste von sieben Stufen.

Diese Überlegung führt uns zu einem weiteren Nachteil introspektiver Urteile, nämlich dass sie subjektiv sind. Laut dieser Kritik ist jede einzelne Urteilsbefragung keine Beobachtung eines externen Ereignisses sondern die Aussprache einer persönlichen Meinung. Solche Daten lassen sich aber nicht unabhängig messen, also können sie nicht angefochten werden. Damit ist das wissenschaftliche Prinzip der Überprüfbarkeit und Replizierbarkeit von Feststellungen nicht gewährleistet.

Das dritte große Problem introspektiver Daten können wir als die Unsicherheit des gemessenen Konstrukts zusammenfassen. Urteile abzugeben ist weder produktive noch rezeptive Verarbeitung, sondern es scheint sich um eine unmotivierte Metakompetenz zu handeln. Es ist also nicht sofort eindeutig, dass gerade dieses Messinstrument als definierendes Kriterium taugt, ob eine Struktur zur Sprache gehört oder nicht. Vorkommensdaten liefern viel eindeutigere Ergebnisse: wenn Sprecher eine Struktur benutzen, muss sie zur Sprache gehören.

Zudem ist es trotz eingehender Studien (siehe Schütze 1996) noch unklar, welche Faktoren in welchem Verhältnis und mit welcher Interaktionsfunktion bei der Bildung eines Gesamturteils beteiligt sind. Fest steht nur, dass u. a.

viele syntaktisch irrelevante Faktoren involviert sind, wie zum Beispiel die Plausibilität des Inhalts, der Bekanntheitsgrad des benutzten Wortschatzes und ähnliches. Damit lässt sich nicht auseinanderhalten, welche Teile eines Urteiles tatsächlich theoretisch relevant sind und welche nicht. Dies muss als zusätzliches Hindernis gelten, wenn man die empfundene Wohlgeformtheit und nicht das Vorkommen als Kriterium für Grammatikalität anwenden will.

Experimentell erhobene Urteile

Diese immer wieder gegen introspektive Urteile vorgebrachten Vorwürfe sind sachlich richtig, betreffen aber nur die Urteile, die mit der Standardmethode der Selbstbefragung erhoben werden. Dagegen ist es aber möglich und aus empirischer Sicht sogar notwendig, seine Daten unter strengerer Kontrolle zu erheben.

Das erste Gebot der empirischen Adäquatheit ist es, keine einzelnen Urteile zu sammeln. Dazu bieten sich mehrere Vorgehensweisen an, von denen wir hier nur drei erwähnen werden. Die herkömmlichste Art, eine solche Studie auszuführen, ist wohl die Benutzung einer Fünf- (oder Sieben)-Punkte-Skala (z. B. Crain/Fodor 1987). Der Versuchsleiter gibt dem Informanten eine graphische Skala mit einer normalerweise ungeraden Zahl von darauf gekennzeichneten Intervallen (Abbildung 1).

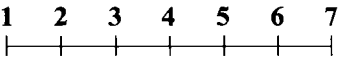

Abb. 1: Eine Sieben-Punkt-Skala zur Erhebung von relativen Urteilen

In letzter Zeit ist dagegen die Methode *Magnitude Estimation* (Bard et al. 1996) beliebter geworden. Drei Unterschiede zur normalen Urteilserhebung prägen diese Methode. Erstens werden Urteile numerisch abgeben. Zweitens werden nur relative Urteile erhoben: relativ zu einem Referenzsatz und zu den eigenen vorherigen Urteilen. Drittens ist die Skala ohne Endpunkte und ohne Mindestabstand. Die Aufgabe lautet dementsprechend: ‚Wenn der Referenzsatz zehn wert ist, wieviel geben Sie diesem Beispiel?'. Diese Methode erlaubt es dem Informanten, alle Wohlgeformtheitsunterschiede auszudrücken, die er oder sie empfindet, ohne von einer vorgeschriebenen Skala eingeschränkt zu sein.

Die von uns bevorzugte Methode nennt sich *Thermometer-Urteile* (Featherston 2004). Dieser Ansatz unterscheidet sich von Magnitude Estimation darin, dass statt eines einzigen Referenzsatzes immer zwei angegeben werden, deren Referenzwerte vom Versuchsleiter festgesetzt werden. Der Grund für diese Neuerung ist, dass sich herausgestellt hat, dass Informanten gar keine proportionalen Urteile abgeben können. Sprecher haben gar keine Intuition, ob eine Struktur ‚doppelt' oder ‚halb' so gut ist wie eine andere, wie das Magnitude Estimation verlangt. Tatsächlich haben wir nur ein Gespür für die

Distanz zwischen den Wohlgeformtheitsgraden von einzelnen Strukturen, und Informanten benutzen in der Praxis eine lineare Skala, auch wenn sie dazu aufgefordert sind, eine proportionale Skala zu benutzen (Featherston 2004, Poulton 1989).

Für die Zwecke unseres Papiers müssen wir jedoch nicht scharf zwischen den Ergebnissen dieser drei Methoden unterscheiden. Das Wichtigste ist, *dass* ein experimenteller Ansatz verwendet wird, und nicht so sehr *welcher*. Tatsächlich liefern diese drei Methoden sehr ähnliche Ergebnisse, was die Robustheit der Befunde unterstreicht. Diese Methoden haben vier wichtige Gemeinsamkeiten:

- 25+ Informanten
- 10+ Lexikalisierungen
- Urteile in numerischer Form
- kontrolliertes linguistisches Experimentmaterial.

Wenn diese vier Anforderungen erfüllt sind, dann verlieren die Kritikpunkte gegen introspektive Urteile weitgehend an Kraft. Weshalb das so ist, werden wir im Folgenden erläutern.

Wenden wir uns wieder den Vorwürfen zu. Zuerst haben wir beanstandet, dass Urteile als Datentyp ungenau sind. Sie sind variabel über Informanten und über Zeit, sie sind unquantifiziert, und sie sind nicht statistisch erfassbar. Dies alles trifft natürlich auf die einzelnen Urteile eines alleine in seinem Büro arbeitenden Syntaktikers zu, nicht jedoch auf diese Resultate von Urteilsexperimenten. Dadurch dass die Urteile numerisch erhoben werden, sind die Schwankungen erfassbar. Mit Werkzeugen wie Mittelwert und Standardabweichung lassen sich empirisch gestützte Aussagen machen. Resultatsmuster wie in Abbildung 2 zeigen, dass Urteile einer gewissen Fehlervarianz unterliegen, daher die berechtigte Kritik bezüglich der Schwankung. Abbildung 2 belegt aber auch eindeutig, dass die Fehlervarianz nicht regellos ist.

Diese Graphik stellt ein typisches Resultat einer Studie mit relativen Urteilen dar. Die getesteten syntaktischen Bedingungen werden auf der horizontalen Achse angeordnet, die vertikale Dimension quantifiziert empfundene Wohlgeformtheit, wobei höhere Zahlen eine größere Natürlichkeit der getesteten Sätze bedeuten. Die Fehlerbalken zeigen für die jeweilige Bedingung ein 95% Konfidenzintervall für den Mittelwert.[1] Die Urteile sind *z*-Werte, d. h. relativ zum individuellen Mittelwert und zur individuellen Standardabweichung in eine normalisierte Skala überführt, damit die Urteile aller Informanten sinnvoll graphisch dargestellt werden können.

Diese Graphik zeigt, dass die Schwankung der Urteile nicht beliebig sondern mehr oder weniger normalverteilt ist. Manche individuellen Urteile bewegen sich weiter weg vom Mittelwert, aber sie bleiben trotzdem in relativer

[1] Diese Graphik zeigt die Ergebnisse einer Studie zur Diskursgebundenheit (*discourse linking*) im Deutschen (Featherston 2005a).

Nähe dazu. Angenommen eine Bedingung ist auf einer Sieben-Punkte-Skala konsensuell eine Zwei-Wert, so fallen neben den vielen Zweien auch Einser- und Dreier-Urteile an, eventuell auch ein Vierer-Wert, aber Fünfer-, Sechser-, und Siebener-Urteile kommen nicht vor. Das heißt, und darum geht es in dieser Diskussion, dass einzelne Urteile vom Mittelwert abweichen mögen, aber im Schnitt sind sich alle Sprecher einig. Dieser Effekt wird aber erst dann sichtbar, wenn man eine genügend große Stichprobe von Sprechern befragt. Aus dieser Perspektive kann man auch gut verstehen, weswegen manche Linguisten die Zuverlässigkeit von Urteilen angezweifelt haben: stellt man zwei Informanten die Frage, wie sie eine gewisse Struktur einschätzen, kann es sehr gut sein, dass der erste mäßig nach oben variiert, während der zweite etwas nach unten tendiert. Die zwei Urteile scheinen disjunkt zu sein. Dieser Tatbestand erlaubt unter anderem folgende Schlüsse: dass Urteile schwankungsanfällig und daher unbrauchbar sind, oder dass die zwei Informanten unterschiedliche Idiogrammatiken haben. Beide Folgerungen werden manchmal gezogen, aber sie sind keineswegs notwendig, denn die Erhebung weiterer Daten lässt das wahre Bild der Fehlervarianz erscheinen. Sie wird durch die üblichen Maße des Mittelwerts und der Standardabweichung erfasst und kontrolliert.

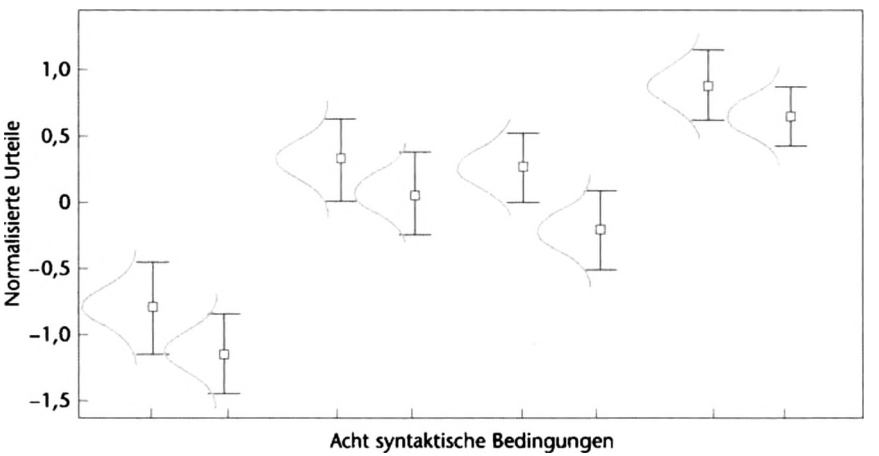

Abb. 2: Die Resultate einer Studie mit relativen Urteilen auf einer Intervallskala. Fehlervarianz verteilt sich normal um einen Mittelwert.

Es dürfte klar sein, dass nicht nur Varianz innerhalb einzelner Bedingungen sondern auch zwischen den Bedingungen quantifiziert werden kann. Bei Urteilen auf einer Intervallskala kann man sich nützlicher statistischer Verfahren wie der Varianzanalyse bedienen. Damit kann man empirisch festhalten, welche Unterschiede signifikant sind.

Auch der Vorwurf, dass Urteile als Datentyp subjektiv sind, ist nicht haltbar, wenn sie systematisch erhoben werden. Die Befragung von mehreren Informanten macht Urteile automatisch zumindest intersubjektiv, denn die

Mittelwerte sind gerade durch die Effekte bestimmt, die in der Informantengruppe überwiegen. Das Auftreten eines Effekts in einer Gruppe von 25 Informanten erlaubt darüber hinaus ganz klar die Vorhersage, dass ein ähnlicher Effekt in einer weiteren Stichprobe von Sprechern zu finden wäre. Damit sind die Effekte externalisiert und objektiv, und sie lassen sich daher auch replizieren. Jedenfalls sind die Ergebnisse insoweit objektiv, als dass sie nicht von der Willkür und Voreingenommenheit des Syntaktikers abhängen können.

Einen weiteren Schritt zur Objektivierung dieses Datentyps kann man mit der Einbeziehung von Standardsätzen als Füllsätzen tun. Für das Deutsche haben wir Gruppen von jeweils fünf Sätzen ermittelt und zusammengestellt, die die gesamte Skala der empfundenen syntaktischen Wohlgeformtheit abdecken. Wenn man in jedem Experiment eine oder zwei solcher Gruppen als Füllsätze einbindet, hat man eine Vergleichsgrundlage. Diese erlaubt es, die Werte aus verschiedenen Experimenten sinnvoll miteinander zu vergleichen und sie würde, wenn die Standardbeispiele besser bekannt wären, Urteilswerte nahezu absolut werden lassen. Die Auswirkungen dieser Technik werden klarer, wenn man sich den Nutzwert einer anderen geläufigen Standardskala überlegt. Wie viel schwieriger wäre es, von Temperaturen zu sprechen, wenn es nicht die Celsius-Skala gäbe! Tatsächlich ist die Wahl des Gefrier- und Kochpunkts von Wasser eher willkürlich (siehe Fahrenheitskala, Kelvinskala), aber der Effekt ist deutlich: ein bekanntes Vergleichsmaß macht die Vorstellbarkeit und Kommunizierbarkeit von Temperaturen viel einfacher.

In Abbildung 3 sieht man anhand eines Beispiels, wie Standardbeispiele eingesetzt werden. In dieser Studie wurden die Dritte Konstruktion und das Lange Passiv miteinander verglichen, und zwar über eine Hierarchie von Verben hinweg. Die Fragestellung betraf die Reaktion dieser zwei Konstruktionen auf die lexikalischen Merkmale des Matrixprädikats (Featherston, unveröffentlicht). Der Befund: die Werte für die Dritte Konstruktion sind immer besser als die für das Lange Passiv, sie werden aber dennoch als sehr unnatürlich eingeschätzt, wie man im Vergleich zu den Standardbeispielen sieht. In so einem Fall kann eine Konstruktion zwar ‚besser' sein als eine andere, aber sie ist deshalb nicht automatisch Teil der Sprache.

Die systematische Erhebung von Urteilen erlaubt es auch, den Vorwurf zu kontern, dass das gemessene Konstrukt bei Urteilen nicht klar definiert sei. Bei diesem Vorgehen lassen sich Einschränkungen präzisieren, was gemessen wird und was nicht. Dadurch dass wir mehrere Informanten befragen, können wir behaupten, dass wir eine Generalisierung über Sprecher messen. Dadurch dass wir mehrere Lexikalisierungen testen, können wir behaupten, dass wir eine Generalisierung über die Struktur messen. Es ist auch Teil der experimentellen Methode, verschiedene irrelevante Störfaktoren auszuschließen oder zumindest zu reduzieren. Wir kontrollieren im Experimentmaterial die Form (lexikalische Häufigkeit, Wortlänge usw), den Inhalt (Plausibilität), den

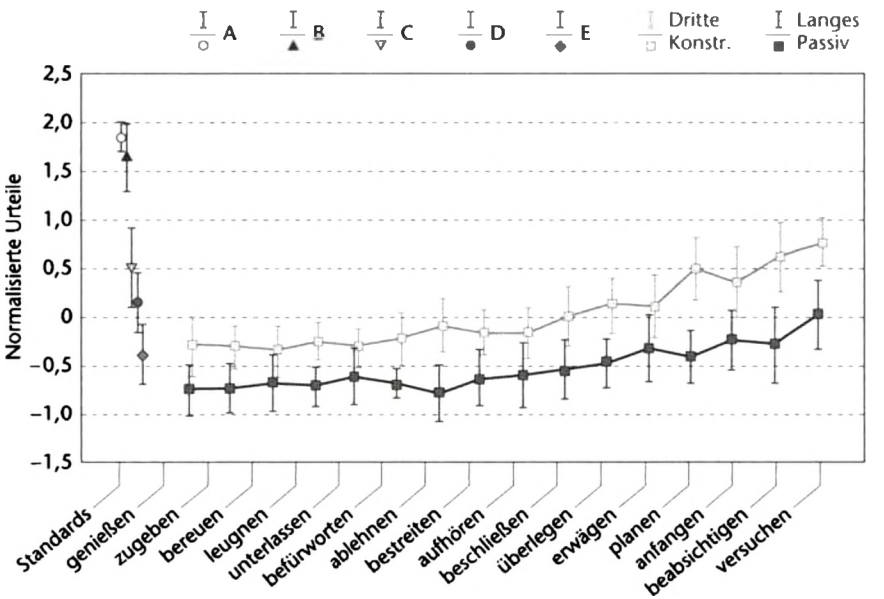

Abb. 3: In dieser Studie wurde eine Gruppe von Standardbeispielen als Vergleichsmaß benutzt. Diese fünf bekannten Werte sieht man auf der linken Seite der Graphik. Die Werte für die Dritte Konstruktion sind immer besser als diejenigen für das Lange Passiv, sie werden aber im Vergleich zu den Standardbeispielen als sehr unnatürlich eingeschätzt.

Kontext (man kann entweder Kontext variieren und explizit angeben oder einfach weglassen, dann nehmen die Informanten den zugänglichsten Kontext), und den Kommunikationsbedarf (in einem Experiment abwesend). Dadurch entfallen viele Störfaktoren.

Es ist interessant, Chomskys (1965, S. 3) Präzision des Anwendungsgebiets der linguistischen Theorie mit den tatsächlichen Merkmalen experimentell erhobener Urteilsdaten zu vergleichen.

> Linguistic theory is concerned primarily with an ideal speaker-listener in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by memory limitations, distractions, shifts of attention and interest, and errors (random and characteristic) in applying his knowledge in actual performance.

Aus diesem berühmten Zitat können wir entnehmen, dass Chomsky über individuelle und lokale Unterschiede, über Verarbeitungseffekte (das heißt zeitdruckabhängige Effekte), über inhalts- und kommunikationsbezogene Faktoren sowie über kontextbezogene Faktoren hinwegidealisieren will. Das System an sich ist von Interesse, und nicht die Art, wie das System in konkreten Situationen angewendet wird. Nun entspricht diese Wunschliste weitgehend dem, was man mit experimenteller Kontrolle erreichen kann. Es ist selbstverständlich nicht der Fall, dass experimentelle Urteile Zugang zur

reinen Kompetenz erlauben, es bleibt noch viel Rauschen in den Daten, aber näher als mit experimentellen Urteilen kommt man beim jetzigen Stand der Technik nicht an Chomskys Definition. Schon allein dadurch sollten Urteile für Linguisten von Interesse sein.

Zusammenfassend stellen wir fest, dass experimentell erhobene, relative Urteile quantifizierbar, statistisch erfassbar, intersubjektiv, replizierbar, und verhältnismäßig frei von Störfaktoren sein können. Zudem haben sie zwei große Vorteile: sie erlauben eine höchst fokussierte Erhebung und trotzdem liefern sie eine sehr feine Differenzierung. Man kann genau die Daten sammeln, die man für einzelne Punkte der Theoriebildung benötigt, und sei die Struktur noch so obskur oder selten. Und trotz der Seltenheit des Phänomens bekommt der Linguist klare, empirisch einwandfreie und statistisch signifikante Ergebnisse. Das macht diese Methode zu einem sehr leistungsstarken Werkzeug für die Syntaxforschung.

Ein letzter Kommentar: gerade die Robustheit der Ergebnisse ermöglicht es, diese Methode nur teilweise anzuwenden, wenn relativ einfache Resultate genügen. Das bedeutet, dass man oft keine allzu aufwändige experimentelle Vorgehensweise benutzen muss. Individuelle Intuitionen von Wohlgeformtheit korrelieren mit Experimentergebnissen sehr stark: wenn ein Resultat für eine Einzelperson nicht nachvollziehbar ist, dann ist das ein Zeichen dafür, dass dieses Resultat nur mit größter Vorsicht zu behandeln ist. Die Experimente liefern zwar mehr Details als die Intuitionen eines einzelnen Sprechers, aber sie sind letzten Endes der gleiche Datentyp, den man nur mit einer höheren Auflösung beobachtet. Dies wiederum bedeutet, dass die Urteile einer einzelnen Person nicht unterbewertet werden sollten, allerdings darf diese Person nicht voreingenommen sein. Den Urteilen eines einzelnen Linguisten kann man und sollte man nur geringen Glauben schenken, aber wenn fünf Unbeteiligte das gleiche behaupten, dann ist das mit hoher Wahrscheinlichkeit ein ernstzunehmender Befund. Anders ausgedrückt: ein verhältnismäßig großer Anteil der begründeten Zweifel an introspektiven Urteilen hängt mit der unsystematischen und unkontrollierten Art ihrer Erhebung zusammen. Sie haben jedoch keine intrinsisch mindere Qualität als Evidenz.

Einblicke in die Syntaxtheorie

In diesem zweiten Teil werden wir darlegen, welches Bild von der Syntax wir mit experimentell erhobenen Urteilen erhalten. Obwohl diese Daten viele Eigenschaften der Standardmodelle bestätigen, werden einige wichtige Annahmen falsifiziert, was für unser Bild von der Architektur der Grammatik Erneuerungen erzwingt. Ich werde im Folgenden für zwei Thesen argumentieren, die sowohl für die Syntaxtheorie selbst wie auch für den Evidenzwert von Korpusdaten von Bedeutung sind.

Die Ergebnisse von Urteilsstudien sehen so aus wie in Abbildung 4. Wie in den vorigen Graphiken zeigt die y-Skala die empfundene Wohlgeformtheit

an, die die Informanten numerisch ausgedrückt haben. Die Werte für jede Bedingung werden mit einem Fehlerbalken angegeben, aus dem der Mittelwert und das 95% Konfidenz-Intervall abzulesen sind. In diesem Beispiel haben wir acht syntaktische Bedingungen, die aus je drei binären Parameter bestehen. Die Fehlerbalken der einzelnen Minimalpaare sind mit einer Linie verbunden. Auf der *x*-Achse haben wir die Werte für die drei binären Parameter angegeben. Wenn die Beschränkung erfüllt ist, bekommt der Parameter den Wert 1, wenn sie verletzt wird, bekommt der Parameter den Wert 0.

Wenn wir die Resultate betrachten, so stellen wir zunächst fest, dass die Beurteilungen nicht, wie ein binäres Modell der Wohlgeformtheit es vorhersagen würde, in zwei Gruppen erscheinen (grammatisch und ungrammatisch), sondern ein Kontinuum darstellen. Auch in keiner anderen Studie ließ sich eine Binarität der Urteile nachweisen. Anstelle von zwei Gruppen sehen wir, dass die Bewertungen direkt auf die Anzahl und Schwere der verletzten Beschränkungen reagieren. Die Verletzung einer Beschränkung hat eine konstante Auswirkung auf die Bewertung einer Struktur. Dies erkennen wir, wenn wir die Werte für Minimalpaare betrachten. Das Verhältnis zwischen den Minimalpaaren, die eine bestimmte Beschränkung erfüllen und verletzen, ist konstant. Das heißt, die Verletzung einer einzelnen bestimmten linguistischen Beschränkung hat immer die gleiche Auswirkung, unabhängig davon, wie viele Beschränkungen eine Struktur sonst noch verletzt. Diesen Befund sehen wir immer wieder: Informanten benutzen kein binäres Modell der Wohlgeformtheit, wenn es ihnen nicht aufgezwungen wird. Man sollte im Auge behalten, dass dieser Befund nicht durch Performanzfaktoren wegerklärt werden kann, da diese in einer experimentellen Umgebung gerade kontrolliert werden.

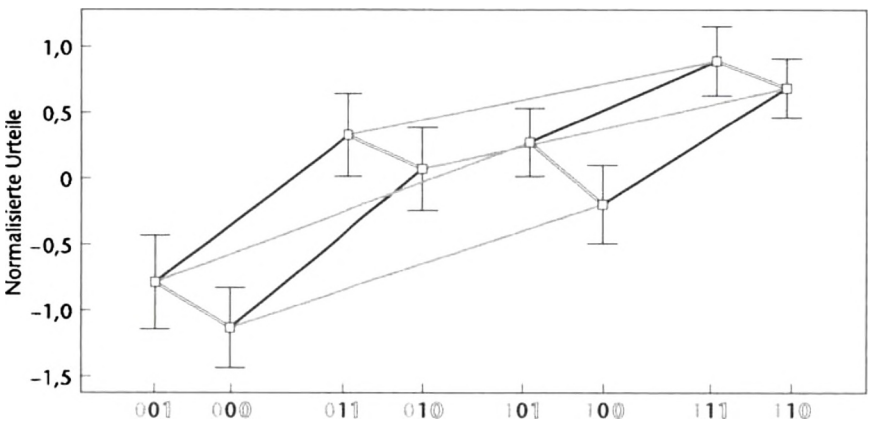

Abb. 4: Diese Graphik zeigt exemplarisch, wie die Resultate von Experimenten mit relativen Urteilen aussehen. Diese Ergebnisse zeigen Wohlgeformtheit als Kontinuum und quantifizierbare, kumulative Beschränkungsverletzungskosten. Es gibt kein Anzeichen für ‚harte', kategorische Beschränkungen.

Anhand solcher Daten müssen wir konstatieren, dass Beschränkungen nicht kategorisch zu sein scheinen. Eine kategorische Beschränkung könnte sich bei diesem Datentyp in zwei möglichen Verhaltensweisen zeigen. Die eine Möglichkeit wäre, dass die verletzende Struktur so schlecht wird, dass sie durch zusätzliche Verletzungen nicht noch schlechter gemacht werden kann. Das würde dem Zustand der absoluten Unwohlgeformtheit entsprechen, denn was ausgeschlossen ist, kann nicht weiter bestraft werden. Diesen Zustand findet man jedoch nie. Strukturen können immer noch schlechter gemacht werden, denn das Bild der strukturellen Wohlgeformtheit, das wir von diesen Daten bekommen, verhält sich kumulativ (Keller 2000).[2]

Eine zweite Möglichkeit, eine ‚harte' Beschränkung zu erkennen, wäre, dass jede sie verletzende Struktur auf ein gleiches Niveau fällt, egal wie gut sie ohne diese Verletzung eingeschätzt wird. Wenn das eintreten würde, wäre die Quantifizierbarkeit der Verletzungskosten nicht mehr gegeben, und die Annahme wäre begründet, dass diese Beschränkung eine verletzende Struktur kategorisch als Teil der Sprache ausschließt. Aber auch dies kommt nicht vor. Die Kosten in Form von verminderter Wohlgeformtheit sind bei einer bestimmten Verletzung immer gleich und deshalb quantifizierbar. Man kann Beschränkungen mit stärkeren und andere Beschränkungen mit schwächeren Verletzungskosten erkennen, wie auch hier in der Abbildung, aber unser Datentyp unterstützt nicht die Annahme, dass es kategorische Beschränkungen in der Sprache gibt.

Bis jetzt haben wir gesehen, dass Informanten, wenn sie die freie Wahl der Skala haben, keine harten Beschränkungen erkennen, sondern immer nur quantifizierte, kumulative Verluste in der empfundenen Wohlgeformtheit. Dies erzwingt ein Bild der Wohlgeformtheit als Kontinuum. Wir sollten jedoch noch eine weitere Feststellung notieren, nämlich, dass wir kein Anzeichen von Verletzbarkeit im Sinne der *violability* der Optimalitätstheorie (OT, Prince/Smolensky 1993) ausmachen können. In der OT haben Beschränkungen immer nur dann eine Auswirkung, wenn sie zwischen den übrigen strukturellen Kandidaten differenziert. EVAL kann nicht zählen, nicht einmal bis eins, was bedeutet, dass diese Funktion zwischen dem Fall, bei dem alle noch aktuellen Kandidaten eine Beschränkung verletzen, und dem Fall, bei dem keiner der aktuellen Kandidaten eine Beschränkung verletzt, nicht unterscheidet. In keinem dieser beiden Fälle hat die Beschränkung eine Auswirkung auf die laufende Evaluation. Das bedeutet aber, dass die Beschränkung keine Anwendung findet. Diesen Tatbestand findet sich jedoch nicht in unseren Urteilen: alle Beschränkungen werden immer angewendet, alle Verletzungen haben immer eine Auswirkung auf die Urteile. Es spielt für eine Beschränkung keine Rolle, ob eine ansonsten perfekte Struktur vorliegt,

---

[2] Dies betrifft nur *strukturelle* Wohlgeformtheit, wohl gemerkt. Wenn Informanten einen Satz nicht mehr verstehen, wird das Bild verschwommener. Aber in unseren Studien geht es um die syntaktische Form, nicht um den Inhalt von Strukturen.

oder ob die Struktur bereits aus unabhängigen Gründen schlecht ist: die Beschränkung selbst findet immer Anwendung. Dementsprechend stützt dieser Datentyp auch nicht die Konzepte der probabilistischen oder optionalen Anwendung von Beschränkungen. Im Gegenteil deuten diese Daten darauf hin, dass die Anwendung von Beschränkungen blind und automatisch erfolgt, wie dies in der generativen Syntax traditionell angenommen wurde.

Diese Befunde sind in relativen Urteilen konstant, robust und unvermeidbar. Aber unsere Schlussfolgerungen für die Natur der Wohlgeformtheit stimmen nicht mit den gängigen Annahmen überein, die ebenfalls aufgrund von Urteilen entwickelt worden sind. Wie kann es sein, dass introspektive Urteile zwei so widersprüchliche Bilder abgeben können? Der Grund liegt darin, dass wir zwei Sorten von Urteilen unterscheiden müssen: relative und kategorische.

Relative Urteile und kategorische Urteile

In unseren empirischen Studien haben wir herausgefunden, dass relative Urteile, die wir mit Sorgfalt unter strenger Kontrolle erheben, ganz andere Merkmale aufweisen als traditionell angenommen wurde. Gleichzeitig aber gibt es verhältnismäßig starke Evidenz, dass die traditionell binären Urteile auch psychologisch reell sind. Jeder Sprecher kann mit dem Konzept ‚absolute Grammatikalität' etwas anfangen. Tatsächlich entspricht dieser Ausdruck einer intuitiv vorhandenen Kategorie, die wir haben, obwohl diese in unseren relativen Urteilen überhaupt nicht sichtbar ist. Daher sind wir gezwungen, zwei Typen von Urteilen zu unterscheiden.

Relative Urteile, die quantifizierbare, kumulative, beschränkungsspezifische Verletzungskosten auf einer Kontinuumsskala aufweisen, geben den Komputationsaufwand der Struktur wieder. Mit Komputationsaufwand meinen wir so etwas wie die Umwandlung von ungeformten Botschaften in Wortsequenzen (cf. Culicover/Nowak 2003 und Referenzen dort). Kategorische Urteile dagegen sind eine Aussage, ob eine Struktur gut genug ist, um vorzukommen. Unsere Hypothese ist es deshalb, dass diese zwei Urteilstypen auf unterschiedliche Faktoren reagieren. Relative Urteile quantifizieren den Denkaufwand, den eine Struktur verursacht. Dass wir auf die Menge kognitiver Arbeitsleistung empfindlich reagieren, und dass dies auch dem Bewusstsein zugänglich sein kann, ist aus anderen Kontexten klar; schließlich wundert sich niemand wenn wir wissen, ob eine Rechnung ‚leicht' oder ‚schwer' ist. Es ist entspannender, einen Krimi zu lesen als ein Werk schöngeistiger Literatur, weil ersteres weniger Denkmühe verursacht; wir sind uns dessen auch bewusst.

Kategorische Urteile dagegen sind ein ganz anderes Maß, denn sie drücken die Meinung aus, ob eine Struktur vorkommen würde. Dabei kann man sich zwei Evidenzquellen vorstellen, aufgrund deren diese Aussage gemacht werden könnte: das interne Korpus und der Vergleich mit anderen Kandidaten. Im ersten Fall ist das Vorgehen eher direkt. Der Informant sucht sein

internes Korpus daraufhin ab, ob er Formen dieser Struktur (oder Strukturverbindung) gehört hat. Falls ja, ist die Struktur grammatisch.

Der zweiten Fall ist komplexer und verlangt zuerst eine Erklärung, was wir unter Äußerungsauswahl (*output selection*) verstehen. Man muss davon ausgehen, dass der Sprecher im Produktionsmodus zwischen verschiedenen strukturellen Varianten auswählen muss, welche er produzieren wird. Diese Auswahl wird manchmal zwischen gleich guten Alternativen sein, manchmal zwischen besseren und schlechteren Varianten. In diesen Fällen gehen wir davon aus, dass der Sprecher normalerweise die ‚bessere' auswählt. Die Auswahl kann lexikalisch sein (*Mach die Tür auf!* vs. *Öffne die Tür!*), strukturell (*Er sagt, dass er mich liebt* vs. *Er sagt, er liebt mich*) oder auch thematisch (*Ich belade den Wagen mit Heu* vs. *Ich lade das Heu auf den Wagen*), aber fest steht, dass wir diese Fähigkeit haben, zwischen Alternativen von Äußerungen auszuwählen. Wenn wir ein binäres Urteil abgeben, kann es sein, dass wir mit diesem Auswahlverfahren die zu bewertende Struktur daraufhin abschätzen, ob man nicht eine andere Version bevorzugt hätte. Falls nicht, dann ist die Struktur ‚gut genug, um vorzukommen'.

Es ist vielleicht bemerkenswert, dass Syntaktiker manchmal eine zugespitzte Version dieses zweiten Verfahrens einsetzen, die einen string-technischen Begriff der Grammatikalität zu Grunde legt. Sie scannen einen Satz, um festzustellen, ob ein bekannter Verletzungstyp darin vorkommt. Wenn nicht, dann geben sie der Struktur das Prädikat ‚grammatisch'. Tatsächlich bildet dieses Vorgehen einen dritten Urteilstyp, dessen Wohlgeformtheitsdefinition etwa ‚diese Struktur enthält keine Verletzung, die normalerweise mit Vorkommen inkompatibel ist' lauten könnte.

Wir haben nun drei verschiedene Arten von Urteilen differenziert, die alle in der Linguistik benutzt werden. Unsere Hypothese ist: Relative Urteile reagieren auf die komputationelle Denklast, während kategorische Urteile aussagen, ob eine Struktur vorkommen würde. Die dritte, stringtechnische Urteilsart sagt aus, ob eine bekannte schwerwiegende Beschränkungsverletzung in einem Satz identifiziert werden kann. Soweit zu Urteilsdaten. Nun werden wir das Verhältnis zwischen unseren bevorzugten relativen Urteilen und Vorkommensfrequenzen erläutern.

Relative Urteile und Häufigkeitsdaten

Ein häufiger Befund in der Syntaxforschung ist, dass die Ergebnisse von Studien mit Urteilen und solchen mit Korpusfrequenzen übereinstimmen, wenn es darum geht, die beste Form aus einer Reihe von Kandidaten zu identifizieren (Featherston 2004, 2005b, Kempen/Harbusch 2005). Diese zwei Datentypen liefern jedoch konstrastierende Bilder, wenn man das Verhalten der schwächeren Kandidaten betrachtet. In Frequenzdaten kommen die Verlierer kaum oder gar nicht vor. In relativen Urteilen dagegen werden Strukturvarianten, die zu schlecht sind, um jemals natürlich vorzukommen, genau

so differenziert und gestaffelt bewertet wie vorkommende Strukturen. Unter der Annahme unserer Hypothese, dass relative Urteile eine Bewertung von komputationeller Komplexität darstellen, ist dies auch erklärbar. Korpusdaten können notwendigerweise immer nur Effekte bei Strukturen messen, die auch vorkommen. Frequenzen von Strukturen, die allesamt nicht vorkommen, werden nicht unterschieden. Der Denkaufwand einer noch so schlechten Struktur wird dagegen sehr wohl gemessen, und fließt daher auch in unsere relativen Urteile ein. Das bedeutet, dass relative Urteile Informationen enthalten, die in Frequenzen nie erscheinen können.

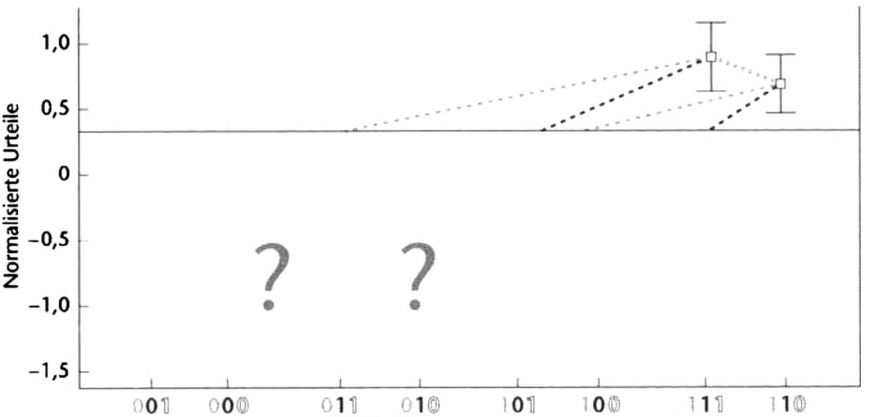

Abb. 5: Die Auswirkung des Eisbergeffekts: Frequenzen können uns immer nur ein Bild von guten Strukturen geben. Was nicht gut genug ist, um vorzukommen, ist unsichtbar. Vergleiche mit dem Bild der relativen Urteile in Abbildung 4.

Wir werden uns in diesem Kontext einer Metapher bedienen. Sprachliche Daten sind wie ein Eisberg: es gibt Strukturen die vorkommen: sie ähneln dem sichtbaren Teil des Eisbergs, der aus dem Wasser ragt. Es gibt noch viel mehr Strukturen, die nie vorkommen: sie bilden den versteckten Teil des Eisbergs unter der Wasseroberfläche. Diese Strukturen sind anhand von Frequenzdaten nicht zu erkennen; man kann lediglich feststellen, dass sie nicht vorkommen.

In diesem Kontext beschränken wir uns darauf, nur ein Ergebnis zu erwähnen, das bei Urteilsdaten zu ungrammatischen Strukturen anders ausfällt als bei Frequenzdaten. Bei Frequenzdaten ist inhärent der Eindruck gegeben, dass die Verletzung einer Beschränkung das Vorkommen der verletzenden Struktur verhindert. Diese Idee des kategorischen Ausschlusses durch Verletzungskosten ist auch natürlich, denn Frequenzdaten bestehen aus vielen individuellen Einzelentscheidungen, ob eine bestimmte Struktur benutzt wird oder nicht. Daher sind die Bausteine von Korpusdaten binär, obwohl das Gesamtbild dieser kategorischen Einzelentscheidungen oft gradierte Formen annimmt. Wenn man mit Frequenzdaten arbeitet, ist es also eine sehr natürliche

Annahme, dass linguistische Beschränkungen kategorischer Natur sind. In Abbildung 5 sieht man skizzenhaft das Bild der Datenlage, das man mit Frequenzdaten bekommt: das obere Glied des Minimalpaares ist sichtbar, das untere oft nicht. Um die Verletzungskosten einer Beschränkung zu quantifizieren, müsste die Distanz zwischen den Elementen eines Minimalpaares sichtbar sein. Man kann die Verletzungskosten einer einzelnen Beschränkung mit Frequenzdaten also nicht messen, wenn das untere Glied eines Paares fehlt. Dies ist immer dann gegeben, wenn eine Verletzung in der Praxis bedeutet, dass eine verletzende Struktur effektiv nicht vorkommt. Dieser Tatbestand vermittelt den Eindruck, dass Strukturen entweder gut sind, oder gar nicht vorkommen (siehe Abbildung 5).

Diese Annahme wird aber durch relative Urteile widerlegt. Tatsächlich sehen die Verhältnisse zwischen Strukturvarianten, die nicht vorkommen, genau so aus wie die zwischen belegten Strukturvarianten oberhalb der Wasseroberfläche. Beschränkungen, die sowohl oberhalb wie auch unterhalb zur Anwendung kommen, benehmen sich überall gleich. Auch die Beschränkungen, die normalerweise als kategorisch eingeschätzt werden, haben quantifizierbare Verletzungskosten. Das legt folgendes Fazit nahe: Obwohl es den Frequenzdaten nach so aussieht, schließen Beschränkungsverletzungen eine Struktur nie automatisch aus. Die Vorkommensgrenze an sich ist für relative Urteile nämlich völlig unsichtbar. Es muss festgestellt werden, dass das Verhältnis zwischen empfundener Wohlgeformtheit, wie sie in relativen Urteilen gemessen wird, und Vorkommen, gemessen anhand von Frequenzen, nicht einfach ist.

Bedeutung für die Architektur der Grammatik

Unsere These ist, dass die Sprachverarbeitungseinheit, die als Grammatik verstanden wird, tatsächlich aus zwei Modulen besteht, die unterschiedliche Funktionsweisen haben. Das erste stellt Kandidaten zur Verfügung, wie die Grammatik dies traditionell tut, und das zweite wählt in einem getrennten Schritt unter den möglichen Strukturen aus.

Sehr vereinfacht dargestellt werden im ersten Modul die verschiedenen linguistischen Beschränkungen auf die Form der Struktur angewendet (siehe Featherston 2005b). Beschränkungen sind nicht gerankt, noch werden sie geordnet angewendet, noch blockieren die einen die Anwendung der anderen, wie die OT dies vorsieht, sondern die Beschränkungen verhalten sich zueinander wie die klassische generative Syntax dies immer vorgesehen hat: blind, simultan, und automatisch. Dieses Modul stellt auch die Verletzungskosten fest: eine Struktur, die eine Beschränkung verletzt, wird um einen beschränkungsspezifischen Wert in ihrer empfundenen Wohlgeformtheit herabgesetzt. Werden zwei oder mehr Beschränkungen verletzt, wird die Struktur noch schlechter, denn diese Verletzungskosten sind kumulativ. Dieser Teil der Grammatik schließt keine Strukturvarianten aus, sondern gibt alle

Varianten mit einer Art Wohlgeformtheitsgewichtung weiter. Wir nennen diesen Teil die Beschränkungsanwendung (*constraint application*). Die Existenz dieses Modul ist notwendig, um die Form der relativen Urteilsdaten zu erklären.

Die zweite Funktion heißt Äußerungsauswahl (*output selection*). Hier wird zwischen den Kandidaten ausgewählt, denn nur eine Form kann tatsächlich produziert werden. Als Auswahlkriterium wird die Wohlgeformtheit berücksichtigt. So kommt es, dass generell nur ‚gute' Strukturen produziert werden. Korpusdaten messen die Ausgabe dieses Moduls, denn sie zählen, was tatsächlich produziert wird.

Beschränkungsanwendung: *constraint application*
- wendet Beschränkungen an,
- stellt Verletzungen fest,
- weist Verletzungskosten zu (Wohlgeformtheitsgewichtung),
- liefert Kandidatenstrukturen mit Gewichtungen zur Auswahl.

Äußerungsauswahl: *output selection*
- wählt aus, welche Strukturvariante produziert wird,
- benutzt Wohlgeformtheitsgewichtungen als Kriterium,
- funktioniert probabilistisch.

Diese Thesen bieten eine Erklärung dafür, weshalb relative Urteile und Frequenzdaten unterschiedliche Ergebnismuster liefern. Die Urteile messen die Wohlgeformtheitsgewichtung aus dem ersten Modul, Frequenzdaten messen das Vorkommen. Da die Konkurrenz um das Vorkommen auf der Basis des Wohlgeformtheitsstatus stattfindet, stimmen Urteilsdaten und Frequenzdaten in ihrer Identifikation der besten Strukturen überein. Suboptimale und schlechte Strukturen dagegen bekommen sehr wohl Wohlgeformtheitswerte, und können deshalb in den Urteilen unterschieden werden.

Diese schwächeren Kandidaten kommen aber in der Praxis nicht vor, denn sie gewinnen nie den Wettbewerb in der Äußerungsauswahl. Sie werden also in den Frequenzdaten nicht unterschieden. Dieser Erklärungsversuch basiert also direkt auf den empirischen Befunden.

Schlussfolgerungen für die Syntaxtheorie

In diesem letzten Teil gehen wir auf zwei Vorteile ein, die sich aus der Unterscheidung der beiden Grammatikmodule ergeben. Der erste betrifft die empirische Adäquatheit der Syntaxtheorie; der zweite die Relevanz der Korpusdaten für die Theorienbildung. Die Syntaxtheorie leidet gerade unter mangelnder Entscheidungsfähigkeit. Nehmen wir die kontroverse Rolle der Konkurrenz in der Theorie als Beispiel. Syntaktiker benutzen mehrere in ihrer Architektur völlig inkompatible Grammatikmodelle. Die *Government and Binding*-Gruppe von Syntaxmodellen (Chomsky 1981) enthält gar keine Kon-

kurrenz. Strukturen sind inhärent entweder gut oder schlecht. Beschränkungen werden blind und automatisch angewendet. Head-driven Phrase Structure Grammar (Pollard/Sag 1994) beinhaltet auch nur ansatzweise eine Konkurrenz-Funktion, in der *obliqueness hierarchy* in Kapitel 6. Hier muss ein Argument weniger oblique sein, um ein anderes binden zu können. Im Minimalistischen Programm (Chomsky 1993) spielt das economy principle eine größere Rolle. Hier konkurrieren Strukturvarianten eindeutig gegeneinander; deren Wohlgeformtheit ist somit relativ zu einer Vergleichsgruppe. In der OT (Prince/Smolensky 1993) findet die Konkurrenz in der Syntax seine Apotheose. In der OT gibt es gar keine absolute Wohlgeformtheit, sondern nur relative. Auch Beschränkungen haben keine sichere Existenz in der Sprache, denn nur diejenige haben eine Auswirkung, die zwischen Kandidaten unterscheiden. Die Anwendung von Beschränkungen ist deshalb nicht blind und automatisch sondern bedingt. Dieser Aufstieg der Konkurrenz in der Syntax wird von Müller/Sternefeld (2001) mit Weitblick beschrieben.

Diese Modelle unterscheiden sich zum Teil recht stark, aber sie werden alle als Erklärungen für den mehr oder weniger gleichen Evidenzbestand angeboten. Wir haben nur ihre Architekturunterschiede in Bezug auf einen einzigen Parameter dargestellt, nämlich den Platz, der dem Konzept der Konkurrenz eingeräumt wird. Dass Syntaktiker unterschiedliche Analysen spezifischer Strukturen bevorzugen, sollte nicht verwundern, dass sie aber so völlig inkompatible Beschreibungsrahmen annehmen können, muss als besorgniserregend gelten.

Auf den ersten Blick lässt dies zwei Erklärungen zu: entweder betrachten Syntaktiker die Evidenz nicht, die zwischen den Modellen entscheiden könnte, oder aber die Evidenzbasis der Syntax ist so dünn, dass man nicht einmal die groben Züge der Architektur der Syntax erkennen kann. Keine der beiden Möglichkeiten gibt ein sehr positives Bild unseres Feldes.

Akzeptiert man aber unser Modell der zwei Grammatik-Module, so eröffnet sich eine neue Perspektive. Das erste Modul, *constraint application*, funktioniert blind, automatisch und kumulativ, ohne jegliche Konkurrenz. Die Ausgabe dieses Modul, so haben wir argumentiert, ist uns durch relative Urteile zugänglich. Diese Funktionsweise entspricht in etwa den Annahmen der *Government and Bindung*-Theorie. Das zweite Modul, *output selection*, ist eine Konkurrenzfunktion; Strukturen konkurrieren aufgrund ihrer Wohlgeformtheit um das Vorkommen. Die Ausgabe dieses Moduls sind Frequenzdaten. Die Funktionsweise entspricht der der Optimalitätstheorie. Es scheint also, dass die starken Abweichungen in den Grammatikarchitekturen jeweils einen Teil der Realität der Funktionsweisen der Grammatik widerspiegeln. So kann es kommen, dass Grammatikmodelle mit und ohne Konkurrenzelemente einigermaßen empirisch adäquat sein können. Korpusdaten beinhalten Beweise dafür, dass die Produktionsweise von Äußerungen unter anderem auch kompetitiv arbeitet. Wenn man Wohlgeformtheit über Vorkommen definiert, dann scheint ein Konstrukt der relativen Wohlgeformtheit

sinnvoll und empirisch begründet. Verwendet ein Linguist eher (nichtexperimentelle) Urteile, bekommt er kein Indiz für Konkurrenz, sondern für die blinde Funktionsweise des ersten Moduls der Beschränkungsanwendung, die absolute Wohlgeformtheit erzeugt.

Wir müssen daher weder die Schlussfolgerung ziehen, dass Syntaktiker das Beweismaterial ignorieren, noch dass die Evidenzlage zu mager ist, um gesicherte Erkenntnisse zu erlauben. Die derart unterschiedlichen Grammatikmodelle können nicht alle richtig sein, aber sie erfassen alle einen Teil des Gesamtbildes. Dieses ist jedoch etwas komplexer als die einzelnen Grammatikmodelle annehmen. Sie treffen auf Schwierigkeiten, weil sie versuchen, die unterschiedlichen Merkmale von zwei unterschiedliche Funktionsweisen auf dieselbe Weise in einem einzigen Modell zu erfassen.

So viel zum ersten Vorteil unserer Unterscheidung von zwei Grammatikmodulen. Der zweite Vorteil betrifft den Evidenzwert von Korpusbelegen. Findet man eine Struktur in einem Korpus, so bedeutet dies, dass diese Struktur im Wettbewerb gegen andere Strukturvarianten gewonnen hat. Nun hat aber dieser Wettbewerb, wie wir auch oben erwähnt haben, einen probabilistischen Charakter, was zur Folge hat, dass manchmal auch die zweitbeste oder sogar die drittbeste Struktur, gemessen an ihrer Wohlgeformtheit, erscheinen kann.

Dieses Merkmal der Arbeitsweise der Äußerungsauswahlfunktion ist durchaus empirisch begründet und anerkannt. Viele korpusbasierte Studien untersuchen, welche Faktoren die Auswahl der einen oder der anderen Strukturvariante bevorzugen (z. B. Bader/Häussler 2006, Bresnan et al. 2005). Dabei ist es aber wesentlich, dass diese Korpusdaten kein kategorisches Gesamtbild zeigen. Nicht Vorkommen und Abwesenheit werden verglichen, sondern höhere und niedrigere Frequenzen. In jedem einzelnen Produktionsfall ist es deshalb nur eine Frage von Wahrscheinlichkeiten, ob der Sprecher zum Beispiel *Den Studenten hat der Vortrag gefallen* oder *Der Vortrag hat den Studenten gefallen* sagt, denn die beitragenden Faktoren halten sich in diesem Fall ungefähr die Waage (Subjekte im Vorfeld werden bevorzugt aber belebte Experiencer im Vorfeld werden auch bevorzugt). Damit ist aber auch vorgegeben, wofür wir argumentieren: die Äußerungsauswahl ist eine probabilistische Kompetition auf der Basis von Wohlgeformtheit.

Wenn dieser Schluss einmal akzeptiert ist, gilt es die Konsequenzen zu untersuchen. Natürlich wird die ‚beste' Strukturvariante am häufigsten vorkommen. Dass auch nur etwas weniger ‚gute' Kandidaten ebenfalls vorkommen, wenn auch mit minderer Frequenz, leitet sich aus der probabilistischen Funktionsweise der Äußerungsauswahl ab.

Aber auch deutlich weniger ‚gute' Strukturen werden zwar selten aber regelmäßig vorkommen, denn die Probabilistik sagt voraus, dass auch unwahrscheinliche Ereignisse hin und wieder stattfinden. Für uns Linguisten bedeutet dies, dass wir in Korpusdaten selten aber vorhersehbar mehr oder weniger ‚schlechte' Strukturen finden werden, ohne dass dieser Befund irgend-

welche Schlussfolgerungen für die Grammatik nach sich ziehen muss. Man kann aus dieser Evidenz nichts über die kausalen Faktoren schließen, die die verschiedenen Beschränkungen im Kern der Grammatik bedingen, denn es ist die probabilistische Auswahlfunktion (output selection), die das Vorkommen dieser ‚schlechten' Beispiele bewirkt. Ihr Vorkommen in einem Korpus ist daher keine Evidenz, dass sie als ‚gut' einzuschätzen sind. Ihr Erscheinen in Korpusdaten lässt uns höchstens schließen, dass sie nicht allzu viel schlechter sind als die Strukturvarianten, die normalerweise vorkommen.

Wenn Linguisten Belege von Strukturen in einem Korpus finden, so nehmen sie dies als Evidenz, dass diese Struktur grammatisch ist. Unsere Hypothese zum bimodularen Charakter der Grammatik stellt dies in Frage. Der erste Teil enthält die Beschränkungen, ihre Anwendung, ihre Interaktion, und die Feststellung der Wohlgeformtheit von Strukturen: somit entspricht er dem Kernteil der Grammatik. Ein gesonderter zweiter Teil wählt zwischen den Varianten aus, und zwar unter Berücksichtigung der im ersten Modul zugewiesenen Wohlgeformtheitsgewichtung. Damit ist gewährleistet, dass allgemein immer nur die besten Strukturen in Korpusdaten erscheinen. Aber da diese Auswahlfunktion probabilistisch arbeitet, können manchmal auch suboptimale Varianten die Auswahlkonkurrenz gewinnen, was bedeutet, dass sie dann in Korpora zu finden sind. Findet der Linguist also wider Erwarten seltene Belege für eine suboptimale Struktur, bedeutet dies nicht, dass er sein Grammatikmodell anpassen muss, um diesem Befund Rechnung zu tragen (kontra z. B. Müller 2003).

Zusammenfassung

In diesem Beitrag haben wir dargelegt, unter welchen Umständen introspektive Urteile objektive, quantifizierbare und empirisch adäquate linguistische Daten sein können. Hierzu ist ein experimenteller Ansatz notwendig. Wenn man eine Vielzahl von Informanten befragt, werden auch introspektive Urteile effektiv objektiv, denn sie lassen Vorhersagen zu, die falsifiziert werden können. Die Erhebung von Urteilen in numerischer Form auf einer Intervallskala erlaubt die Quantifizierung von Urteilen, die Ermittelung von deren Mittelwerten und Standardabweichung, sowie die Anwendung nützlicher statistischer Testverfahren. Wenn man auch verschiedene lexikalische Varianten der syntaktischen Bedingungen testet, dann lässt sich die Behauptung machen, dass die Resultate frei von lexikalischen Effekten sind. Durch den experimentellen Kontext werden viele Störfaktoren ausgeschlossen oder deren Auswirkung abgemildert. Damit erfüllen diese experimentell erhobenen, relativen Urteile die Anforderung der empirischen Adäquatheit soweit das im Rahmen des Möglichen erreicht werden kann. Weil sie eine sehr genau fokussierte Erhebung erlauben und zudem noch klare, robuste Ergebnisse liefern, sind diese Urteile eine sehr attraktive Datenquelle für die Syntax.

Im zweiten Teil haben wir beschrieben, wie die Resultate solcher Studien aussehen. Sie zeigen ein Kontinuum der Wohlgeformtheit, ohne binäre Unterscheidung zwischen ‚gut' und ‚schlecht'. Sie zeigen, dass alle Beschränkungen immer angewendet werden und dass die Verletzung einer Beschränkung eine quantifizierbare, beschränkungsspezifische Reduzierung der empfundenen Wohlgeformtheit zur Folge hat. Die Grenze zwischen Strukturen, die vorkommen, und solchen, die nicht vorkommen, ist dagegen unsichtbar. Daraus schließen wir, dass relative Urteile etwas anderes messen als das Vorkommen.

Um diesen Tatbestand zu erklären, haben wir unsere Hypothese von den zwei Grammatikmodulen entwickelt. Der erste Teil wendet linguistische Beschränkungen an und ermittelt Wohlgeformtheit. Der zweite ist eine Konkurrenzfunktion und wählt die Struktur aus, die produziert wird. Wir haben argumentiert, dass diese Architektur eine Lösung zu zwei Problemen in der Syntaxforschung bietet. Zum ersten kann sie erklären, weshalb Grammatikmodelle mit völlig unterschiedlichen Architekturen empirisch erfolgreich sein können. Die einen spiegeln die Funktionsweise des ersten Moduls wider, die anderen die Funktionsweise des zweiten Moduls. In einem weiteren Schritt haben wir darauf hingewiesen, dass die Architektur voraussagt, dass selten aber regelmäßig sub-optimale Strukturen in Korpusdaten erscheinen werden. Diese Belege erzwingen aber keine Änderungen in der Grammatik, denn sie erscheinen nur wegen des probabilistischen Charakters des Auswahlverfahrens. Nur ein relativ häufiges Vorkommen hat Implikationen für die Wohlgeformtheit.

Bibliographie

Bader, Markus/Häussler, Jana (2006): Word-order variation: Why corpus and judgment data do not go hand in hand! Poster bei der Tagung ‚Linguistic Evidence', Tübingen.

Bard, Ellen/Robertson, Dan/Sorace, Antonella (1996): Magnitude estimation of linguistic acceptability. In: Language 72 (1), S. 32–68.

Bresnan, Joan/Cueni, Anna/Nikitina, Tatiana/Baayen, Harald (2005): Predicting the Dative Alternation. Erscheint in: Royal Netherlands Academy of Science. Workshop on Foundations of Interpretation proceedings.

Chomsky, Noam (1981): Lectures on Government and Binding. The Pisa Lectures. Berlin: Mouton de Gruyter.

Chomsky, Noam (1995): The Minimalist Program. Cambridge, Massachusetts: MIT Press.

Crain, Steven/Fodor, Janet (1987): Sentence matching and overgeneration. In: Cognition 26, S. 123–169.

Culicover, Peter/Nowak, Andrzej (2003): Markedness, antisymmetry and complexity of constructions. In: Pica, Pierre/Rooryk, Johann (Hg.): Variation Yearbook. Amsterdam: Benjamins.

Featherston, Sam (2004): Judgements in syntax: Why they are good, how they can be better. Vortrag bei der DGfS Jahrestagung 2004.

Featherston, Sam (2005a): Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. In: Lingua 115/11, S. 1525–1550.

Featherston, Sam (2005b): The Decathlon Model: Design features for an empirical syntax. In: Reis, Marga/Kepser, Stephan (Hg.): Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton de Gruyter. S. 187–208.

Keller, Frank (2000): Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Disseration, University of Edinburgh.

Kempen, Gerard/Harbusch, Karin (2005): The relationship between grammaticality judgements and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Reis, Marga/Kepser, Stephan (Hg.): Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives. S. 329–350. Berlin: Mouton de Gruyter.

Labov, William (1975): What is a linguistic fact? In: Austerlitz, Robert (Hg.): The Scope of American Linguistics. Lisse: Peter de Ridder. S. 77–133.

Lakoff, George (1973): Fuzzy grammar and the performance/competence terminology game. In: Chicago Linguistics Society 9, S. 271–291.

Müller, Gereon (1995): A-bar Syntax. A Study in Movement Types. Studies in Generative Grammar 42. Berlin/New York: de Gruyter.

Müller, Gereon/Sternefeld, Wolfgang (2001): The rise of competition in syntax: A synopsis. In: Müller, Gereon/Sternefeld, Wolfgang, (Hg.): Competition in Syntax. Berlin: Mouton de Gruyter.

Müller, Stefan (2003): Mehrfache Vorfeldbesetzung. In: Deutsche Sprache 31, S. 29–62.

Pollard, Carl/Sag, Ivan (1994): Head-driven Phrase Structure Grammar. Chicago: University of Chicago Press.

Poulton, Edward (1989): Bias in Quantifying Judgments. Hove & London: Lawrence Erlbaum.

Prince, Alan/Smolensky, Paul (1993): Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science Technical Report 2.

Sampson, Geoffrey (2001): Empirical Linguistics. London/New York: Continuum.

Schütze, Carson (1996): The Empirical Basis of Linguistics. Chicago: University of Chicago Press.

Wurmbrand, Susanne (2001): Infinitives: Restructuring and Clause Structure. Berlin/New York: Mouton de Gruyter.