Zuerst ersch. in: Zeitschrift für Rezensionen zur germanistischen Sprachwissenschaft 2 (2010), 1, S. 25-31, DOI 10.15/zrs.2010.007

Karl-Heinz Best. 2006. *Quantitative Linguistik. Eine Annäherung* (Göttinger Linguistische Abhandlungen 3). 3., stark überarbeitete und ergänzte Auflage. Göttingen: Peust & Gutschmidt. iv, 154 S.

Heike Zinsmeister

Fachbereich Sprachwissenschaft
Universität Konstanz, Fach D 185
D-78457 Konstanz
Heike.Zinsmeister@uni-konstanz.de

Durch die Zipf'schen Gesetze hat die Quantitative Linguistik Eingang in das Allgemeinwissen der Linguistik gefunden. Das prominenteste lautet: Ordnet man die Wörter eines Textes nach ihrer Häufigkeit in Rängen, kann man feststellen, dass relativ wenige Ränge Wörter aufweisen, die sehr häufig sind, wohingegen sich auf sehr vielen Rängen Wörter tummeln, die nur selten belegt sind. Georg Kingsley Zipf hat als erster erkannt, dass es hier einen regelhaften Zusammenhang zwischen der Frequenz eines Wortes und seinem Platz in der Rangordnung gibt. Das Produkt aus Rang r und Frequenz f eines beliebigen Wortes ergibt über alle Ränge nahezu einen konstanten Wert C: $r * f = C$ (Zipf 1949: 24, mit Daten aus Ulysses, nach S. 9). Die Tatsache, dass sich entsprechende Verhältnisse in praktisch jedem Text wiederfinden lassen, erheben die mathematische Formel zu einem Sprachgesetz. Auf der praktischen Seite sagt dieses Zipf'sche Gesetz voraus, dass man in jedem Text auf das Problem stoßen wird, viele Wörter mit nur wenigen Belegen zu finden, egal wie umfangreich die Textbasis ist – eine Tatsache, die weit reichende Konsequenzen für die Auswertung von Texten und Korpora hat (Baroni 2009).

Die Zipf'schen Gesetze bilden nur die Spitze des Eisbergs von mathematischen Gesetzmäßigkeiten in der Sprache. In der Quantitativen Linguistik ist das Entdecken dieser Gesetzmäßigkeiten ein Selbstzweck, in benachbarten Disziplinen könnten diese Gesetze als Referenzgrößen dienen. Da in der Germanistik, anders als in vielen anderen Disziplinen, Statistik nicht zur Standardausbildung gehört, erhalten quantitative Erkenntnisse wenig Aufmerksamkeit und statistische Sprachgesetze werden allenfalls als exotische Kuriositäten wahrgenommen. An diesem Zustand möchte Karl-Heinz Best mit seinem Buch *Quantitative Lin-*

DOI 10.15/zrs.2010.007

guistik. Eine Annährung, welches 2006 nach 2001 und 2003 in einer dritten „stark überarbeiteten und ergänzten" Auflage erschien, etwas ändern.

Das Buch richtet sich an „Philologen (Linguisten, aber auch Literaturwissenschaftler), die nie daran gedacht haben, sich ernsthaft mit sprachstatistischen Phänomenen zu befassen" (S. 5). Seine Ausführungen wenden sich „in erster Linie an die nicht spezialisierten Leserinnen und Leser [...], in der Hoffnung, bei ihnen einiges Interesse wecken zu können" (S. 6), da er „ein offenkundiges Bedürfnis vieler Linguisten und Literaturwissenschaftler an präzisen Aussagen über ihre Beobachtungen [erkennt], [...] z. B. wenn versucht wird, stilistische Besonderheiten eines Textes, einer Textsorte/Gattung oder eines Autors zu beschreiben" (S. 5).

Best möchte einen möglichst komprimierten Überblick geben, hält daher die Vorstellung von einzelnen Untersuchungen kurz und verzichtet auf eine vertiefende Darstellung. Um dem interessierten Leser trotzdem eine weitere Auseinandersetzung zu ermöglichen, macht er sehr ausführliche Literaturangaben. Thematisch konzentriert sich der Autor auf die Überprüfung von Gesetzeshypothesen, die wir unten genauer vorstellen werden und legt diese „vorwiegend aus der Perspektive der Quantitativen Linguistik (QL) in den deutschsprachigen Ländern" dar (S. 5). Den ersten Zugang zur Welt der Quantitativen Linguistik erhält der Leser durch eine kurze Einführung in die historische Entwicklung der Disziplin und die Vorstellung ihrer wichtigsten Vertreter, Veranstaltungen und Fachorgane sowie Hinweise auf mehrere aktuelle Forschungsvorhaben im deutschsprachigen Raum. Nach dieser allgemeinen Übersicht grenzt Best fünf inhaltliche Themenbereiche der Quantitativen Linguistik ab (S. 11f.):

i. Statistische Analysen von Wörtern und Texten zur „Lösung ganz konkreter Aufgaben" wie der Entwicklung des Morsealphabets oder der Chiffrierung und Dechiffrierung im Allgemeinen;
ii. Untersuchungen zu „Lösung von Problemen in Nachbardisziplinen" wie z. B. „Forschungen zu Lesbarkeit und Textverständnis für die Psycholinguistik";
iii. Sprachstatistische Erhebungen zur „Differenzierung linguistischer Befunde oder Hypothesen", d. h. der Verifizierung oder Falsifizierung von quantitativen Aussagen in der linguistischen Analyse wie z. B. der Hypothese, dass „hochfrequente Wörter sich dem Sprachwandel widersetzen";
iv. Stilistik, z. B. Untersuchungen zum Nominalstil;
v. Überprüfung von sprachstatistischen Gesetzeshypothesen wie z. B. den Formeln, welche den Zipf'schen Gesetzen zugrunde liegen.

Der fünfte Themenbereich, die Diskussion von verschiedenen Sprachgesetzen, bildet den Schwerpunkt von Bests Buch. Bevor er sich auf diesen Bereich konzentriert, führt der Autor die Leser zunächst in sehr grundlegende Überlegungen zur Wortschatzstatistik ein. Er diskutiert Fragen wie z. B. die Wortschatzgröße des Deutschen und die Ermittlung des aktiven und passiven Wortschatzes einzelner Sprecher. Zur Bestimmung des Gesamtwortschatzes von Kindern berichtet Best zum Beispiel von einem Schätzverfahren, das auf Kenntnissen zum Wortschatzzuwachs in Texten beruht (Wagner, Altmann & Köhler 1987, nach 18ff.). Diese Schätzung soll hier beispielhaft auch für die Überprüfung von Sprachgesetzen in Abschnitt 3.2 vorgestellt werden.

Zerlegt man einen Text in gleich lange Textstücke, z. B. mit je 200 Wörtern, nimmt die Anzahl der neu auftretenden Wörter vom ersten bis zum letzten Textstück regelhaft ab. Diese Abnahme kann mit einer Formel modelliert werden, bei der die berechnete Anzahl der neuen Wörter y eine Funktion der fortlaufenden Textblocknummer x ist: $y=ax^b$, wobei a und b zwei Parameter sind, die für jede Verteilung neu angepasst werden, so dass sich die vorhersagten Werte möglichst nahe an die gemessenen Werte annähern. Die Güte der Anpassung wird durch einen statistischen Test ermittelt (S. 29), bei dem überprüft wird, ob die gemessenen Daten, gegeben ist eine gewisse zufällige Streuung, statistisch gesehen mit der berechneten Verteilung übereinstimmen können. Durch die Formel kann vorausgesagt werden, ab welchem (fiktiven) Textblock mit weniger als einem neuen Wort zu rechnen ist, gleichzeitig kann die bis dahin geäußerte Menge an neuen Wörtern geschätzt werden, welche mit dem Vokabular des Kindes gleich gesetzt wird. Best zitiert als illustratives Beispiel die Untersuchung einer 430 Minuten langen Aufnahme der 12 Jahre und 2 Monate (12,2) alten Christiane. Bei einer Parameterbelegung von a =43,6 und b = -0,4098 ergibt sich für Christiane ein geschätzter Gesamtwortschatz von 47.321 Wörtern.

Nach den einleitenden Kapiteln führt Best im Hauptteil des Buches die Leser anhand einer Vielzahl von Fallbeispielen durch acht Sprachgesetze. Jedem Gesetz liegt eine mathematische Formel bzw. eine Gruppe von Formeln zugrunde. Die Auszählung der Einzeluntersuchungen wird immer in Tabellenform dargestellt und in den meisten Fällen zusätzlich als Balken- oder Kurvendiagramm visualisiert. Die Sprachgesetze lauten im Einzelnen:

I. Die Verteilung von Wortlängen gemessen in Silben, Morphen, Phonemen oder Buchstaben, einschließlich der Anwendung der Theorie auf andere Sprachgrößen wie zum Beispiel Satzgliedtiefe oder die Länge von Illokutionsketten (S. 23-64) lassen sich mit Varianten der Formel $P_x = g(x)P_{x-1}$

beschreiben, wobei P_x die Wahrscheinlichkeit für das Auftreten eines Wortes mit der Wortlänge x ist, P_{x-1} die Wahrscheinlichkeit für ein Wort mit der Wortlänge $x-1$ und $g(x)$ für eine Funktion steht, welche den proportionalen Unterschied zwischen x und $x-1$ bestimmt. Eine prominente Annahme für $g(x)$ lautet $g(x) = a/(b+x)$.

II. Rang-Frequenz-Verteilungen von Buchstaben, Phonemen und Wörtern (S. 73-80) werden nicht wie in der Einleitung angedeutet mit einem der Zipf'schen Gesetze modelliert, sondern mit der so genannten Zipf-Mandelbrot-Verteilung, die eine bessere Annäherung an die Verteilung der sehr häufigen und sehr seltenen Wörter erlaubt.

III. Das Diversifikationsgesetz (S. 80-91) beschreibt die statistische Verteilung von verschiedenen Bedeutungen oder Ausdrucksformen einer Entität, z. B. die Distribution der Pluralallomorphe des Deutschen in einem Text als Ausdrucksformen des nominalen Plurals.

IV. Das Martin'sche Gesetz (S. 92f.) modelliert die Proportionen der verschiedenen Ebenen einer „Definitionskette" in einem Lexikon. Die Ebenen der Kette unterscheiden sich durch den Grad der Verallgemeinerung, z. B. „Sessel – Sitzmöbel – Möbel – Einrichtungsgegenstand – Gegenstand", und durch die Anzahl der Ausdrücke, welche die einzelnen Ebenen belegen. Die Veränderung in der Ausdrucksanzahl in Bezug auf die Ebenen wird durch das Gesetz vorhergesagt.

V. Das Menzerath-Altmann-Gesetz erfasst den Zusammenhang von der Größe eines sprachlichen Konstrukts und der Größe seiner Konstituenten (S. 94-98). Es besagt, dass je größer ein sprachliches Konstrukt ist, desto kleiner sind seine Konstituenten. Beispiele hierfür sind die Abnahme der Morphlänge, der Silbenlänge oder der Polysemie mit zunehmender Wortlänge.

VI. Das Gesetz von Zwirner, Zwirner und Frumkina zur Verteilung von Wörtern in Segmenten eines Textes oder auch das „Textblockgesetz" (S. 99ff.) modelliert die Verteilung von einzelnen Wörtern in fortlaufenden Segmenten eines Textes.

VII. Das Gesetz der Wortschatzdynamik (S. 104ff.) wurde bereits zur Ermittlung des Wortschatzes bei Kindern eingeführt. Es greift auf die einfachste Formel des Menzerath-Altmann-Gesetzes zurück und modelliert den Umfang des Wortschatzes einzelner Sprecher oder Werke.

VIII. Das logistische (Piotrowski-)Gesetz zum Sprachwandel (S. 106-123) erlaubt, verschiedene Formen des Sprachwandels

zu modellieren: den vollständigen Sprachwandel, bei dem alte Formen vollständig durch neue ersetzt werden, wie z. B. der diachrone Wandel der 2. Person Singular Indikativ Präsens des Verbs *wollen* von *wilt* zu *willst*, den unvollständigen Sprachwandel, bei dem die Ersetzung nur bis zu einem gewissen Maß stattfindet, z. B. die Übernahme von griechischen Fremdwörtern ins Deutsche, und den reversiblen Sprachwandel, bei dem neue Formen aufkommen und dann wieder ganz oder teilweise verschwinden, wie z. B. die *e*-Epenthese bei Verben: *ware* als Form des Verbs *sein* ist wieder verschwunden. Das Modell für einen unvollständigen Sprachwandel wurde auch zur Modellierung von Wortschatzverlust und von Spracherwerb eingesetzt.

In zwei Exkursen erläutert Best die Handhabung eines der von ihm verwendeten Software-Pakete, Altmann-Fitters (1997/2005), welches eine Vielzahl von Verteilungen bereitstellt und die benötigten statistischen Tests zur Optimierung der Parameter durchführt.

In einem abschließenden Kapitel fasst Best die Sprachgesetze zusammen und umreißt ein Funktionsmodell der Sprache, in das sich die einzelnen Gesetzmäßigkeiten integrieren lassen. Darüber hinaus gibt er einen Ausblick darauf, wie sich ein System aller Gesetzeshypothesen im Sinne einer Theorie der linguistischen Synergetik (nach Köhler 1986) entwickeln sollte und skizziert Perspektiven für die zukünftige Forschung im Bereich der Quantitativen Linguistik. Das Buch schließt mit einem umfangreichen Literaturregister sowie Angaben der verwendeten Software und einem Hinweis auf die Internetpräsenz des *Göttinger Projekts Quantitative Linguistik* (http://wwwuser.gwdg.de/~kbest/).

Es stellt sich die Frage, ob es Best gelingt, mit seinem Buch das Zielpublikum der Philologen zu erreichen, welche bei fachspezifischen Veröffentlichungen aus dem Bereich der Quantitativen Linguistik in Anbetracht der Zahlenkolonnen und komplexen Formelgleichungen wahrscheinlich sofort weiterblättern würden. Sofern sich die Leser auf die einleitenden Kapitel einlassen, bietet der Hauptteil über die Sprachgesetze tatsächlich eine gute Gelegenheit in ein fremdes Fachgebiet hinein zu schnuppern. Beim ersten Sprachgesetz führt Best den Leser über drei Gedankenexperimente und fünf Beispieltexte langsam zu den mathematischen Gesetzmäßigkeiten. Unterstützt wird das Verständnis durch den ersten Exkurs zur mathematischen Software, deren Handhabung und die Interpretation der Ergebnisse ebenfalls Schritt für Schritt vorgestellt werden. Im weiteren Verlauf des Hauptteils zu den Sprachgesetzen gibt Best jeweils die Basisfunktionen an, verzichtet aber weitestgehend auf die Angabe von komplexen Formeln. Die angeführten Literaturangaben ermöglichen es dem interessierten Leser bei

Bedarf auch die mathematischen Details nachzuvollziehen. Um einen möglichst kompakten Überblick geben zu können, reduziert Best leider auch die Diskussion um die Definitionen der zu untersuchenden linguistischen Entitäten, welche für die Philologen unmittelbar nachvollziehbar wären. Wird bei den Modellierungen zum Wortschatz z. B. kritisch hinterfragt, was man als Worteinheit betrachten kann (S. 21), bleibt bei anderen Untersuchungen nur der Rückgriff auf die zitierte Literatur, z. B. in Bezug auf Wortarten (S. 40, 60) oder Wortbildungstypen (S. 85).

Die Datengrundlagen und die daraus geschlossenen Verallgemeinerungen überraschen zum Teil. Kurze Privatbriefe werden als ideale Textgrundlage angegeben. Texte von mehr als 2000 Wörtern seien hingegen problematisch (S. 39). Hierzu hätte man gerne im Buch weitere Erläuterungen erhalten.

Am Ende bleibt ein leichtes Unbehagen, da es für den Laien nicht nachvollziehbar ist, warum sich eine bestimmte Verteilung für ein konkretes Problem besser eignet als eine andere. Ebenso ist es schwer abzuschätzen, ob eine gegebene Annäherung, die den statistischen Test erfüllt, tatsächlich gut genug ist, und welche Verteilung man wählen sollte, wenn mehrere die statistischen Tests erfüllen. Auf Seite 112 hat sich ein Tippfehler eingeschlichen, in der Tabelle wird der Determinationskoeffizient D mit 99.94 angegeben. Er kann aber nur Werte zwischen 0 und 1 annehmen. Kleine Änderungen im Format könnten zur Verständlichkeit beitragen, z. B. werden alle Zahlen gemäß dem englischen Format mit einem Punkt an der Nullstelle angegeben, die Tausenderstelle wird aber nicht abgesetzt. Kapitelnummern würden dem Leser die Orientierung erleichtern. Für ein Buch, das sich insbesondere an Fachfremde richtet, ist ein Stichwortindex eigentlich unabdingbar. Übersichtslisten zu den Formeln, Tabellen und Schaubildern wären zusätzlich hilfreich.

Best gelingt es, einen Eindruck von der Quantitativen Linguistik zu vermitteln und die Bandbreite der Sprachgesetze vorzustellen. Er bietet damit fachfremden Philologen, aber auch Fachleuten, einen Ausgangspunkt zur weiteren Beschäftigung mit diesem Thema und für den Einsatz von Erkenntnissen der Quantitativen Linguistik bei philologischen Fragestellungen. Kleine Änderungen im Format und die Ergänzung eines Stichwortregisters würden das Buch dem fachfremden Zielpublikum leichter erschließen.

Literatur

Altmann-Fitter. 1997/2005. *Iterative Fitting of Probability Distributions*. Lüdenscheid: RAM-Verlag.

Baroni, Marco. 2009. Distributions in Text. In: Anke Lüdeling & Merja Kytö (Hg.). *Corpus Linguistics. An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29). Berlin: Mouton de Gruyter. 803-822.

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Wagner, Klaus, Gabriel Altmann & Reinhard Köhler. 1987. *Die Sprechsprache des Kindes.* Teil 1. Düsseldorf: Schwann.

Zipf, Georg Kinsley. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.