Stefan Engelberg (Mannheim)

Quantitative Verteilungen im Wortschatz

Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons

Abstract: Das Lexikon menschlicher Sprachen basiert auf quantitativen Verteilungen, die sich am Zipfschen Gesetz orientieren: Wenige Lexeme werden extrem häufig verwendet und sehr, sehr viele Lexeme sind extrem selten. Auch funktional zusammenhängende Teilwortschätze wie Wörter einer bestimmten Wortart, Verben, die in einem bestimmten Argumentstrukturmuster auftreten, oder Komposita zu einem bestimmten Grundwort zeigen ähnliche Frequenzverteilungen, weisen aber auch jeweils typische Abweichungen von einer Zipfschen Verteilung auf.

Zipfnahe Verteilungen sind charakteristisch für dynamische, selbstorganisierende Systeme, und Veränderungen im Wortschatz oder in Teilwortschätzen sind insofern auf der Basis solcher Verteilungen zu interpretieren. Der Artikel plädiert dafür, lexikologischen Sprachdokumentationen ein dynamisches Lexikonkonzept zugrunde zu legen, in dem die Verteilungscharakteristika als Grundlage der Wortschatzstruktur eine zentrale Rolle spielen.

1 Zur Dokumentation des deutschen Wortschatzes

Die zentrale Aufgabe des Instituts für Deutsche Sprache ist es, „die deutsche Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte wissenschaftlich zu erforschen und zu dokumentieren“.[1] Die Abteilung Lexik des IDS ist dieser Aufgabe insbesondere hinsichtlich ihres dokumentarischen Teils bisher vor allem in Form lexikografischer Arbeiten nachgekommen, die Teile des deutschen Wortschatzes in Form gedruckter oder internetbasierter Wörterbücher beschreiben. Mit ihrer auf das einzelne Lexem fokussierenden Darstellungsweise können Wörterbücher systematische Eigenschaften des Wortschatzes natürlich nur eingeschränkt erfassen. In diesem Aufsatz soll es darum gehen, Aspekte der quantitativen Verteilung des Wortschatzes und seiner daraus erwachsenden Dynamik darzustellen, die in den bisherigen Wortschatzdarstellungen

1 So die Satzung des IDS in ihrer Fassung vom 12.3.2014.

des IDS nur eine sehr eingeschränkte Rolle gespielt haben, für die Struktur des Wortschatzes aber von grundlegender Bedeutung sind und daher in seiner Beschreibung auch nicht fehlen dürfen.

Die Ermittlung von quantitativen Verteilungen lexikalischer Einheiten in Korpora gehört – zum Beispiel in Form von Wortfrequenzbestimmungen oder Kookkurrenzanalysen – durchaus schon seit längerer Zeit zum Handwerkszeug lexikologischer Forschung und lexikografischer Praxis. Darüber hinaus gibt es eine lange Forschungstradition in der quantitativen Linguistik, die elementare quantitative Strukturen im Wortschatz aufzudecken versucht (siehe etwa Altmann et al. 2002). Diese Forschungen haben aber kaum Eingang in die germanistische Wortschatzforschung gefunden, sei sie nun eher philologisch oder lexikontheoretisch ausgerichtet. In dem vorliegenden Aufsatz sollen einige grundlegende quantitative Wortschatzverteilungen anhand verschiedener lexikalischer Phänomene illustriert werden und hinsichtlich ihres Potenzials zur Aufdeckung der Struktur des deutschen Wortschatzes und seiner Dynamik besprochen werden.

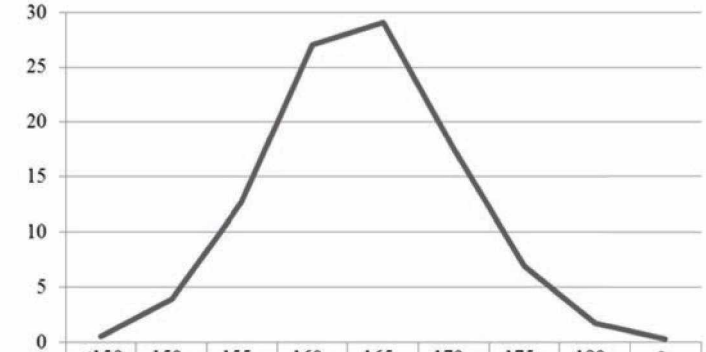
Abschnitt 2 ist grundlegenden Eigenschaften der quantitativen Verteilung des Wortschatzes gewidmet. In Abschnitt 3 werden Besonderheiten in der Verteilung verschiedener funktional zusammenhängender Teilwortschätze dargestellt. In Abschnitt 4 wird es dann darum gehen, zu zeigen, inwiefern spezifische quantitative Verteilungsmuster eine Analyse der synchronen Dynamik des Wortschatzes unterstützen. In der abschließenden Diskussion in Abschnitt 5 wird erörtert, welche Eigenschaften ein als dynamisch verstandenes Lexikon hat und ob traditionelle lexikografische Formate eigentlich geeignet sind, ein solches Lexikon abzubilden.[2]

2 Verteilungen im Wortschatz

Betrachtet man den Bestand und die Häufigkeit von Lexemen in Korpora beliebiger Größe, so ist zu beobachten, dass Lexeme sich extrem in ihrer Gebrauchshäufigkeit unterscheiden. Solche Frequenzunterschiede sind Gegenstand dieses Abschnitts. Dazu sollen zunächst zwei Grundtypen quantitativer Verteilungen illustriert werden, die Gaußsche Verteilung (Normalverteilung) und die Zipfsche Verteilung.

2 Für Kommentare und Diskussionen zu den in diesem Aufsatz dargestellten Phänomenen sei Alexander Koplenig und Irene Rapp gedankt.

Die Gaußsche Verteilung ist dadurch charakterisiert, dass die meisten Werte nahe am arithmetischen Mittel liegen, und das arithmetische Mittel mit dem Median (dem an mittlerer Stelle stehenden Wert, wenn alle Werte der Größe nach sortiert werden) und dem Modalwert (dem am häufigsten auftretenden Wert) zusammenfällt. Die Verteilung spiegelt sich dabei symmetrisch am arithmetischen Mittel, und die Ratio höchster Wert/niedrigster Wert ist üblicherweise sehr viel niedriger als die entsprechende Ratio in den unten zu besprechenden Zipfschen Verteilungen. Typische Beispiele sind die Verteilung von Intelligenzquotienten und die von Körpergrößen. So sind die meisten Menschen ungefähr mittelschlau und mittelgroß.[3] Abbildung 1 zeigt die Verteilung von Körpergrößen bei Frauen. Die gemessene Verteilung ist eine Näherung an die Gaußsche Glockenkurve.



	<150 cm	150–154 cm	155–159 cm	160–164 cm	165–169 cm	170–174 cm	175–179 cm	180–184 cm	≥ 185 cm
Anteil	0,6	4	12,7	27	29,1	17,6	6,9	1,8	0,3

Abb. 1: Prozentualer Anteil von Frauen an Körpergrößenklassen im Jahr 2006[4]

Vergleichbare Verteilungen lassen sich auch bei sprachlichen Daten beobachten: In einem deutschsprachigen Zeitungskorpus (ca. 15 Mio. Textwörter) der Leipziger Korpus Collection kommt das Wort *Mannheim* 438-mal vor. Da die Leipziger Korpora aus zufällig ausgewählten Einzelsätzen aus Online-Zeitungen und ande-

3 Andere Beispiele für Gauß-verteilte Daten sind etwa die Verteilung von gemessenen PKW-Geschwindigkeiten auf einer Landstraße (Newman 2005) oder die Exemplarstreuung innerhalb natürlicher Arten bezüglich körperlicher Merkmale und Verhalten.

4 Die Daten entstammen dem sozioökonomischen Panel. Zitiert nach: Seite „Körpergröße". In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 9. April 2014, 17:51 UTC. Internet: http://de.wikipedia.org/w/index.php?title=K%C3%B6rpergr%C3%B6%C3%9Fe&oldid=129366723 (Stand: 21.4.2014, 11:12 UTC).

ren Internet-Texten bestehen und generell zudem davon auszugehen ist, dass in solchen Texten immer mal wieder zu verschiedenen Anlässen über Mannheim berichtet wird, kann man erwarten, dass *Mannheim* in gleich großen Teilen dieses Korpus auch etwa gleich oft vorkommt. Teilt man nun im Rahmen einer Reliabilitätsprüfung das Korpus sukzessive in 50 gleichgroße Partitionen L01 bis L50 und überprüft das Vorkommen des Wortes *Mannheim* in jeder dieser Partitionen, so ergibt sich folgende Verteilung:

P	f	P	f	P	f	P	f	P	f
L01	8	L11	4	L21	9	L31	6	L41	9
L02	9	L12	10	L22	8	L32	13	L42	9
L03	4	L13	6	L23	9	L33	11	L43	2
L04	10	L14	6	L24	9	L34	6	L44	5
L05	10	L15	10	L25	4	L35	14	L45	10
L06	7	L16	5	L26	10	L36	9	L46	8
L07	8	L17	12	L27	8	L37	15	L47	7
L08	8	L18	6	L28	13	L38	7	L48	12
L09	10	L19	18	L29	14	L39	7	L49	7
L10	10	L20	12	L30	6	L40	10	L50	8

Tab. 1: Häufigkeit (f) des Lexems *Mannheim* in 50 Partitionen (P) des deutschsprachigen Zeitungskorpus aus der Leipziger Korpus Collection

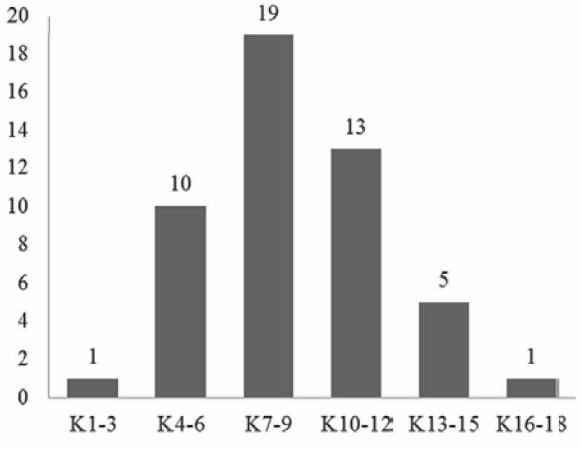


Abb. 2: Zuordnung der 50 Korpuspartitionen zu Häufigkeitsklassen; die x-Achse zeigt die Häufigkeitsklassen (K1–3: Partitionen, in denen *Mannheim* 1–3mal vorkommt, K 4–6: Partitionen, in den *Mannheim* 4–6mal vorkommt, etc.), die y-Achse die Anzahl von Partitionen in jeder Klasse (Spektraldarstellung)

Wenn man sich die Verteilung in Tabelle 1 anschaut, so zeigt sich, dass die meisten Partitionen eine *Mannheim*-Frequenz nahe am Mittelwert aufweisen. In Abbildung 2 sind die Partitionen Klassen zugeordnet, die die Häufigkeiten des Vorkommens von *Mannheim* widerspiegeln. Die Abbildung zeigt eine Näherung an die Gaußsche Glockenkurve. Median (= 9), Modus (= 10) und Mittelwert (= 8,76) sind fast identisch, und die Kurve ist annähernd symmetrisch.

Bei dieser Sichtweise auf die Daten werden auf der x-Achse die Häufigkeitsklassen präsentiert und auf der y-Achse die Anzahl der Partitionen in den jeweiligen Klassen. Dies ist die Spektraldarstellung der Daten. Wir können die Daten aber auch auf andere Weise darstellen. So wie wir Personen wie die Orgelpfeifen nach Größe geordnet nebeneinanderstellen können, lassen sich auch die einzelnen Korpuspartitionen, geordnet nach der Häufigkeit der *Mannheim*-Vorkommen, auf der x-Achse nebeneinander präsentieren, so wie in Abbildung 3. Auf der y-Achse wird dann die Häufigkeit des Wortes Mannheim in den einzelnen Partitionen abgetragen.

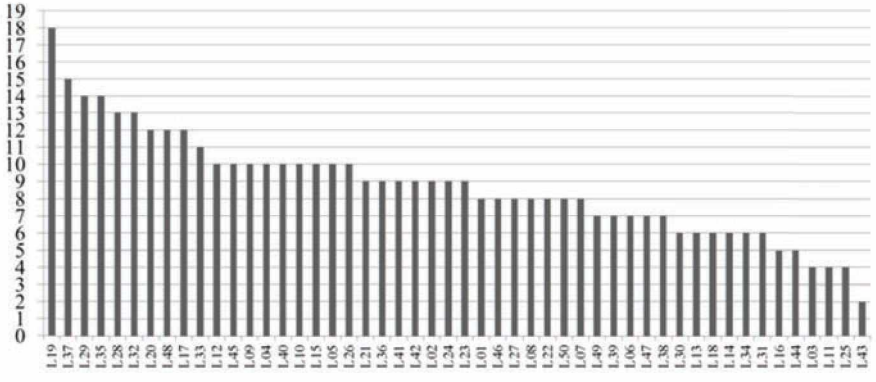


Abb. 3: Korpuspartitionen geordnet nach der Häufigkeit des Wortes *Mannheim* (Rangdarstellung)

Diese Orgelpfeifenanordnung der Daten wird als Rangdarstellung bezeichnet.[5] Im Falle einer Gauß-Verteilung sehen wir hier nicht die typische Glockenkurve, sondern eine Kurve, die einem – meist nur schwach gekrümmten – auf den Rücken gefallenen S ähnelt.

Über solche Distributionsphänomene wie das gerade geschilderte hinaus sind für den Wortschatz allerdings bekanntlich nicht Gaußsche, sondern Zipf-

5 Zur „Spektralverteilung“ und „Rangverteilung“ vgl. auch Tuldava (1998, S. 36).

sche Verteilungen typisch. Zipfsche Verteilungen sind dadurch charakterisiert, dass einige wenige Einheiten sehr häufig auftreten und extrem viele Einheiten äußerst selten. In einer Zipfschen Verteilung ist das zweithäufigste Element halb so häufig wie das häufigste, das dritthäufigste ein Drittel so häufig wie das häufigste usw. (Zipf 1949, S. 22 ff.). Weist man Wörtern gemäß ihrer Häufigkeit einen Häufigkeitsrang r zu, so ergibt sich die Häufigkeit eines Wortes an Rang r wie folgt, wobei α eine (im gegebenen Fall empirisch zu ermittelnde) Konstante ist, die ungefähr 1 beträgt:

Zipf's Gesetz: $f(r) = 1 / r^{\alpha}$

Das heißt, die Auftretenswahrscheinlichkeit eines Wortes ist umgekehrt proportional zu seinem Häufigkeitsrang. Die Verteilung ist nicht symmetrisch; der Mittelwert ist dem niedrigsten Wert deutlich näher als dem höchsten; Median und Modalwert liegen noch einmal deutlich unter dem Mittelwert. Angewandt auf Wortverteilungen in Korpora mittlerer Größe, ist der Modalwert gewöhnlich 1, der Median 1 oder 2.

Auch für zipfnahe Verteilungen kann man die Spektraldarstellung und die Rangdarstellung verwenden: Zur Illustration sollen hier die Frequenzen der Wortformtypen in Kafkas Erzählung „Der Heizer" dienen. Der Text besteht aus 10.014 Textwörtern, die 2.652 Wortformtypen zugeordnet sind. Die Frequenzen der Wortformtypen liegen zwischen 1 und 300 (*der*); der Mittelwert beträgt 3,81, Median und Modalwert sind jeweils 1. In der Rangdarstellung stellt sich die Verteilung wie in Abbildung 4 dar.

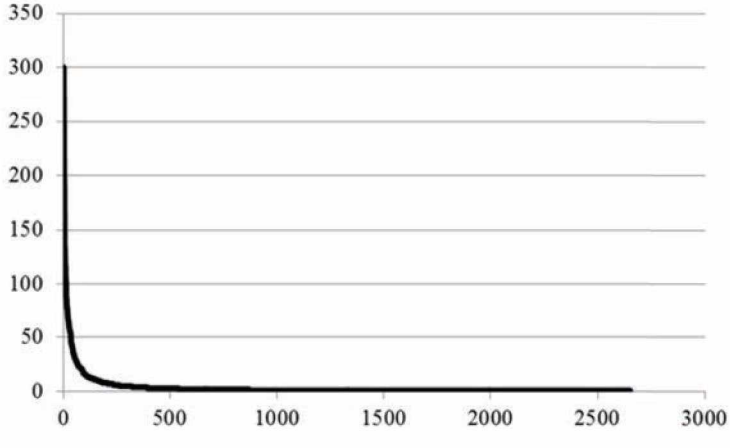


Abb. 4: Häufigkeitsverteilung der 2.652 Wortformtypen in Kafkas Erzählung „Der Heizer"; y-Achse: Wortformfrequenz, x-Achse: Häufigkeitsrang der Wortformen (Rangdarstellung)

Auch hier können wir Häufigkeitsklassen bilden, um die Daten in der Spektraldarstellung zu präsentieren. Teilen wir das Frequenzspektrum (0–300) in Zehnerschritten auf, so stellt sich die Verteilung wie in Abbildung 5 dar.[6]

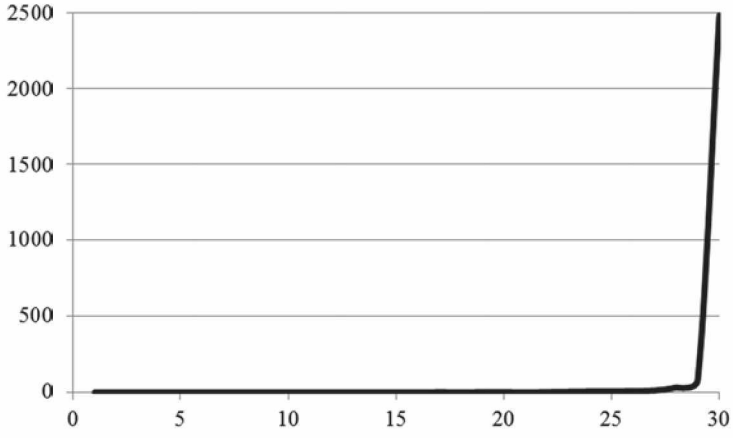


Abb. 5: Häufigkeitsverteilung der 2.652 Wortformtypen in Kafkas Erzählung „Der Heizer" (Spektraldarstellung); auf der x-Achse sind die 30 Größen-Klassen (in 10er Schritten) repräsentiert, auf der y-Achse die Anzahl der Wortformtypen in den einzelnen Klassen

Die Spektraldarstellung zeigt bei absteigender Ordnung der Größenklassen eine extrem linksschiefe, die Rangdarstellung eine extrem rechtsschiefe Verteilung. Beide Darstellungen reflektieren die extreme Ungleichverteilung der Wortformtypen. Von den 2.652 Wortformtypen treten 2.481 (93,55%) nur 1–10-mal auf; 1.775 (= 66,93%) kommen nur einmal vor.

In Abbildung 6 kann man in typisierter Form sehen, wie sich die Gaußsche und die Zipfsche Verteilung in den beiden Darstellungsformen zueinander verhalten. Die Differenzierung in die beiden Darstellungen ist insofern relevant, als Gaußsche Verteilungen meist in der Spektraldarstellung dargeboten werden, in der ja die typische Glockenkurve der Normalverteilung zum Ausdruck kommt, während Zipfsche Verteilungen häufiger in der Rangdarstellung ihren Ausdruck finden.

6 Die Darstellung dient hier nur zur Illustration einer Spektraldarstellung; wie Häufigkeitsklassen sinnvollerweise für lexikologische Zwecke zu konstruieren sind, wird etwa in Perkuhn/Keibel/Kupietz (2012) beschrieben.

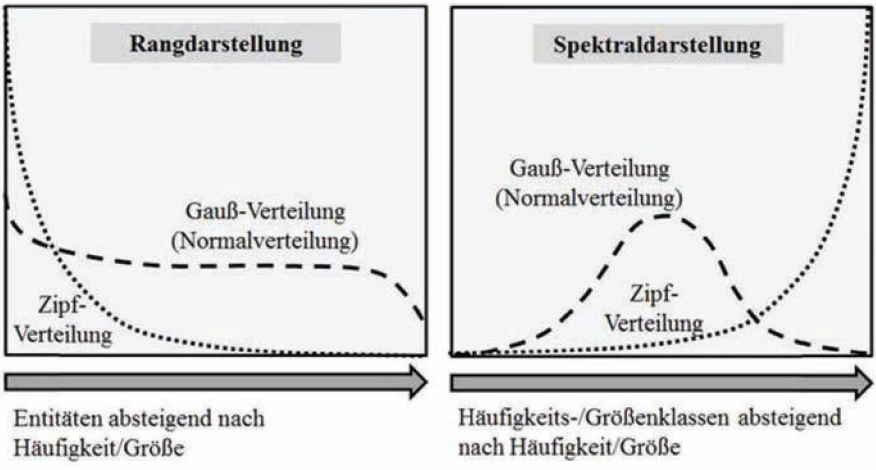


Abb. 6: Vergleich von typisierten Gaußschen und Zipfschen Verteilungen in der Rang- und der Spektraldarstellung

So dramatisch die Abbildungen 4 und 5 die Verteilungen auch darstellen, so wenig lassen sie doch an interessanten Details erkennen, die einen Vergleich mit anderen empirisch ermittelten Verteilungen erlauben würden. Um dies zu ermöglichen, soll zunächst ein Vergleich mit einer idealen Zipf-Verteilung Übereinstimmungen mit und Abweichungen von der Zipfschen Funktion erkennen lassen. Die ideale Zipf-Verteilung für eine gegebene Typen- und Tokenanzahl berechnet sich über die harmonische Reihe, wobei r der Rang ist, LN der natürliche Logarithmus, *TYP* die Anzahl an Typen und *TOK* die Anzahl an Token:[7]

$$\text{Normierungsfaktor für Zipfsche Verteilung} = (1 / (r \cdot \text{LN}(1{,}78 \cdot TYP)) \cdot TOK)$$

Dies sei hier wieder illustriert am Beispiel von Kafkas Erzählung „Der Heizer". In Abbildung 7 sind für die 50 häufigsten Wortformen in Kafkas Text (*die*, *der*, *und*, *ist*, ..., *alles*, *des*, *dieser*, *dort*) die tatsächliche und die Zipfsche Verteilung zu sehen.

Die Darstellung zeigt, dass etwa ab Rang 10 die Wortformen sehr nahe an der erwarteten Verteilung liegen, die Wortformen auf niedrigeren Rängen aber deutlich weniger häufig vorkommen als bei einer Zipfschen Verteilung zu erwarten wäre.

7 Siehe dazu etwa Baayen (2001, S. 15 f.).

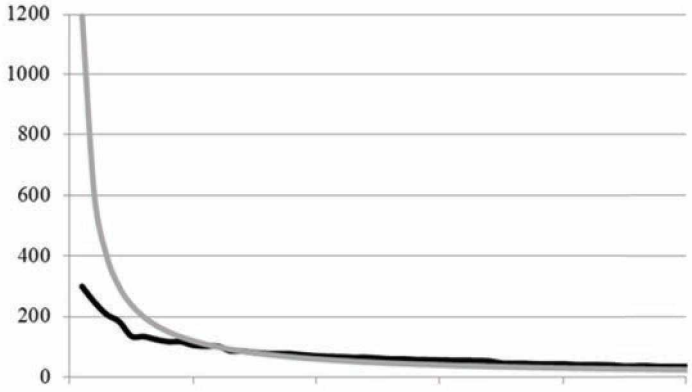


Abb. 7: Häufigkeitsverteilung der 50 häufigsten Wortformtypen in Kafkas Erzählung „Der Heizer" (Rangdarstellung); schwarze Linie: tatsächliche Verteilung; graue Line: ideale Zipf-Verteilung

Bei der Darstellung der Verteilung des gesamten Typenbestands wie in Abbildung 4 entsteht bei großer Typenanzahl das Problem, dass zipfnahe Verteilungen immer mehr oder weniger als rechter Winkel erscheinen. Details sind hier kaum zu erkennen. Für Verteilungen mit vielen Datenpunkten werden daher doppellogarithmische Darstellungen der Verteilungen verwendet, d.h. auf x- und y-Achse werden der Logarithmus der Häufigkeit und der Logarithmus des Wortrangs dargestellt. Die Zipfsche Verteilung stellt sich nun als eine Gerade dar und die tatsächliche Verteilung als eine mehr oder weniger große Annäherung an diese Gerade. Abbildung 8 illustriert das für den „Heizer".

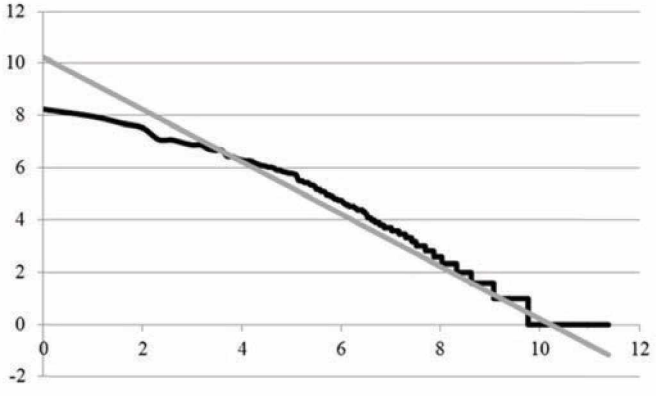


Abb. 8: Häufigkeitsverteilung der Wortformtypen in Kafkas Erzählung „Der Heizer" (doppellogarithmische Rangdarstellung); schwarze Linie: tatsächliche Verteilung; graue Line: ideale Zipf-Verteilung

In der logarithmisierten Darstellung lassen sich nun die für Wortschatzverteilungen typischen Abweichungen der tatsächlichen von der Zipfschen Verteilung gut beobachten: Bei den hochfrequenten Wortformen liegt die tatsächliche Frequenz deutlich niedriger als in einer Zipfschen Verteilung, im mittleren Bereich dagegen etwas höher. Zur besseren Erfassung dieser empirischen Abweichungen von der Zipfschen Verteilung sind eine Reihe von Varianten des Zipfschen Gesetzes vorgeschlagen worden, etwa die Zipf-Mandelbrot-Verteilung[8] (siehe etwa Tuldava 1998, S. 61 ff.) oder das Modified Power Law (Naranan/Balasubrahmanyan 2005, S. 718).

Wie an den obigen Beispielen erkennbar, sind die tatsächlich gefundenen Verteilungen natürlich immer nur Annäherungen an die Zipfsche Verteilung. Abhängig von verschiedenen Parametern weichen die tatsächlichen Verteilungen zum Teil auch systematisch von einer Zipfschen Verteilung ab (siehe z.B. Ferrer i Cancho 2005, S. 250). Zu diesen Parametern gehören unter anderem die Sprache mit ihrer jeweils charakteristischen Wortbildungsmorphologie, die Textsorte oder auch der Sprechertyp: So zeigen Kinder in verschiedenen Stadien des Spracherwerbs charakteristische Abweichungen, und Schizophrene weisen typische Verteilungsauffälligkeiten abhängig vom Erkrankungstyp auf (Ferrer i Cancho 2005, S. 249).

Zipfnahe Verteilungen finden sich in Korpora aller Größen, wobei sich die Verteilungen mit zunehmender Korpusgröße langsam verändern (vgl. etwa Tuldava 1998, S. 75). So gilt bei ansonsten gleichen Rahmenbedingungen, dass der Anteil an Hapax Legomena in einem kleinen Korpus größer ist als in einem großen Korpus. Typischerweise liegt dieser Anteil bei rund der Hälfte der Lexeme.

Die Zipfsche Verteilung mit ihrer hohen Anzahl geringfrequenter Lexeme wird auch reflektiert in den erheblichen Unterschieden zwischen der Anzahl der in Wörterbüchern lemmatisierten Lexeme und der Anzahl der verschiedenen Lexeme, die sich selbst in Korpora mittlerer Größe finden lassen. Große Wörterbücher des Gegenwartsdeutschen haben um die 200.000 Stichwörter. Klein (2013, S. 19) schätzt den Umfang des zehnbändigen Duden (Duden 1999) ohne Verweislemmata auf ca. 175.000 Stichwörter; der sechsbändige Brockhaus-Wahrig (Wahrig 1980–1984, S. 5) verzeichnet nach eigenen Angaben 220.000 Lemmata. Tabelle 2 zeigt demgegenüber die Anzahl an Lexemen, die sich in kleinen und mittelgroßen Korpora ermitteln lässt, in zwei Zeitungskorpora, „Die Welt“ (= RZ; Rosengren 1972, 1977) und die „Süddeutsche Zeitung“ (= SZ; Evert/Baroni 2005) und in einem „Berichtskorpus“ aus Texten verschiedener Textsorten (= BK, Klein 2013).

---

8 In der Zipf-Mandelbrot-Verteilung wird der Rang r um einen Faktor β verschoben (Mandelbrot 1961):

Zipf-Mandelbrot-Gesetz: $f(\mathbf{r}) = \mathbf{1} / (\mathbf{r} + \beta)^{\alpha}$

Korpus	Textwörter	Wortformtypen	Lexeme
RZ	0,0024 Mrd.	0,146 Mio.	0,102 Mio.
SZ	0,2261 Mrd.	2,585 Mio.	1,809 Mio.
BK	1,0000 Mrd.	7,612 Mio.	5,328 Mio.

Tab. 2: Anzahl der Wortformtypen und der Lexeme in Korpora verschiedener Größe[9]

Man sieht, dass schon in einem eher kleinen Korpus wie SZ mit einer viertel Milliarde Textwörter der Lexembestand etwa zehnmal höher ist als in den größten deutschen Standardwörterbüchern. Das Deutsche Referenzkorpus am IDS hat 2014 knapp 25 Mrd. Textwörter, ist also mehr als 100mal so groß. In welchem Maße sich der Lexembestand mit wachsender Korpusgröße erhöht, ist allerdings bisher nur ansatzweise bekannt.[10]

3 Verteilungen in Teilwortschätzen

Zipfnahe Verteilungen lassen sich aber nicht nur für den Gesamtwortschatz einzelner Korpora feststellen, sondern auch für viele funktional zusammenhängende Teilwortschätze. Abbildung 9 zeigt die Häufigkeitsverteilung der Adjektivlexeme in einem Korpus gesprochener Sprache.[11] Insgesamt weist das Korpus 3.754 Adjektivlexeme mit insgesamt 36.981 Vorkommen auf. Häufigstes Lexem ist *gut* mit 2.688 Vorkommen.

Man sieht hier eine Verteilung, wie sie auch für den Gesamtwortschatz typisch ist, mit Werten im oberen und unteren Frequenzbereich unterhalb der Zipf-Verteilung und im mittleren Bereich leicht darüber.

9 Die Anzahl an Wortformtypen wurde mithilfe entsprechender korpuslinguistischer Werkzeuge ermittelt. Die Lexemanzahl in SZ und BK ergibt sich aufgrund der Annahme, dass die Anzahl an Lexemen ca. 30% unter der der Wortformtypen liegt. Die Annahme entspricht sowohl Kleins (2013) Überprüfung als auch den Zahlenverhältnissen in Rosengrens (1972, S. 1315, 1977, S. 1961) Korpora.

10 Vgl. zu einer Diskussion verschiedener Maße Evert/Baroni (2005) und zu den Problemen bei der Lemmatisierung großer Korpora Stadler (ersch. demn.).

11 FOLK-Korpus, Institut für Deutsche Sprache; Stand: Dezember 2013. Es ist hier zu berücksichtigen, dass bei einigen der hochfrequenten Lexeme nicht zwischen ihrer Funktion als Adjektiv und als Gesprächspartikel unterschieden wurde.

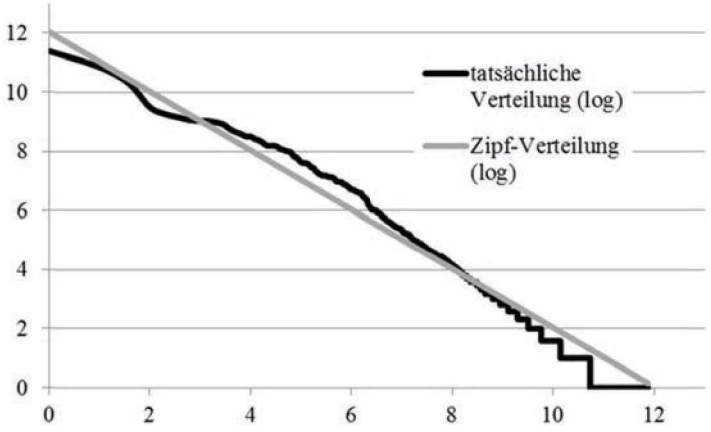


Abb. 9: Verteilung der Adjektive im FOLK-Korpus (doppellogarithmische Rangdarstellung); auf der x-Achse die Rangstellen, auf der y-Achse die Häufigkeiten der Adjektive

Je nach Art des gewählten Wortschatzausschnitts zeigen sich allerdings unterschiedliche Arten von Abweichungen von Zipfschen Verteilungen, die im Folgenden an verschiedenen Beispielen illustriert werden sollen. Kleine relativ geschlossene Wortschatzausschnitte, wie etwa bestimmte Klassen von Funktionswörtern, weisen Verteilungen auf, die den Verteilungen von Graphemen oder Phonemen ähneln, die ja ebenfalls kleine, geschlossene Klassen konstituieren. Der für Zipfsche Verteilungen typische lange Schweif von niedrigfrequenten Einheiten ist hier wenig ausgeprägt. Diese Verteilungen werden nach Naranan/Balasubrahmanyan (2005, S. 729) besser durch das „cumulative modified power law" beschrieben.

Auch die hohe Produktivität von Komposita relativ zu bestimmten Grundwörtern spiegelt sich in charakteristischen Verteilungskurven. Abbildung 10 zeigt die Ergebnisse einer Korpusuntersuchung zur Häufigkeit von Komposita zum Grundwort *gelb*. Eine Suche in DeReKo ergab 30.995 Token, die nach manueller Überarbeitung 1.106 *gelb*-Lexemen zugeordnet werden konnten. Die frequentesten sind *blaugelb* (3.046 Vorkommen), *goldgelb* (2.905) und *rotgelb* (2.221). Um auch einen Eindruck von den Hapax Legomena wiederzugeben, sei hier eine Auswahl aufgelistet: *biedermeiergelb*, *bonbongelb*, *dotterblütengelb*, *dudengelb*, *eiercognacgelb*, *elementargelb*, *emmentalergelb*, *erdnussgelb*, *feuriggelb*, *flippiggelb*, *gehirnhöhlenleitergelb*, *grünblassgelb*, *hellhoniggelb*, *hellviolettsenfgelb*, *inkagoldgelb*, *jasmingelb*, *kalaharigelb*, *knackgelb*, *knuspriggelb*, *limonadengelb*, *löwenzahnblütengelb*, *meranogelb*, *mitrauchergelb*, *nickeldioxingelb*, *oktobergelb*, *ostfriesengelb*, *peachgelb*, *prachtvollgelb*, *quietscheentenknallegelb*, *reineclaudengrüngelb*, *safarigelb*, *sepiagelb*, *softrosadottergelb*, *telefonbuchgelb*, *urgelb*, *vanillepuddinggelb*, *wonnegelb*, *zahnsteingelb*, *zitronenchromgelb*.

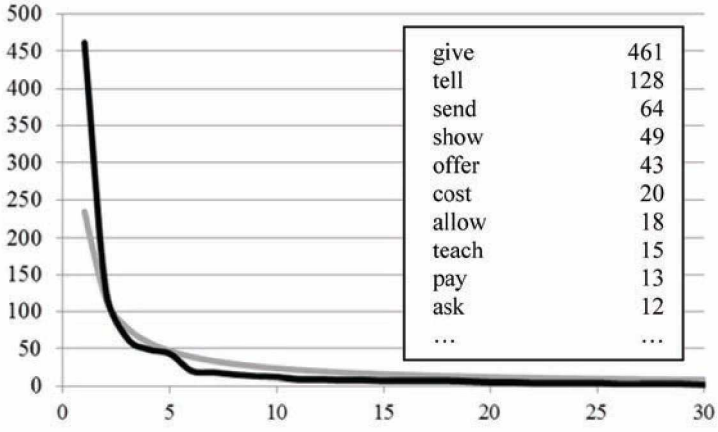


Abb. 10: Verteilung von Komposita, die auf *-gelb* enden (doppellogarithmische Rangdarstellung)

Die Darstellung illustriert einen hohen Anteil an niedrigfrequenten Lexemen, deren Frequenz unter dem Zipfschen Erwartungswert liegt; der Anteil der Hapax legomena beträgt 56,51%. Die hoch- bis mittelfrequenten, vermutlich lexikalisierten Komposita liegen dagegen über den Zipfschen Werten.

Besonderheiten anderer Art zeigen sich dagegen bei Argumentstrukturmustern und den Verben, die in ihnen auftreten. Aus quantitativen Untersuchungen wissen wir, dass Argumentstrukturen Verben in sehr unterschiedlich starkem Maße anziehen (Stefanowitsch/Gries 2003; Engelberg 2015). Dies sei illustriert am Beispiel der Ditransitivkonstruktion im Englischen (Bsp. 1, Abb. 11), des Split-Stimulusmusters mit *an* bei deutschen Psych-Verben (Bsp. 2, Abb. 3) und der direkten und indirekten Redeeinleitung im Deutschen (Bsp. 3, Abb. 13).[12]

(1) a. NP^1 V NP^2 NP^3
b. she gave / told / sent / ... him something

12 Zu den Details der Datenerhebungen zur Ditransitivkonstruktion (basierend auf dem BNC), zum gespaltenen Stimulus (basierend auf DeReKo) und zur Redewiedergabe (basierend auf einem Erzähltextkorpus) vgl. Stefanowitsch/Gries (2003), Engelberg (ersch. demn.) bzw. Jäger (1968). Die Untersuchungen zur Ditransitivkonstruktion und zum gespaltenen Stimulus operieren mit verschiedenen Assoziationsmaßen und nicht mit den absoluten Frequenzen, die hier zur Illustration der Verteilungen herangezogen werden.

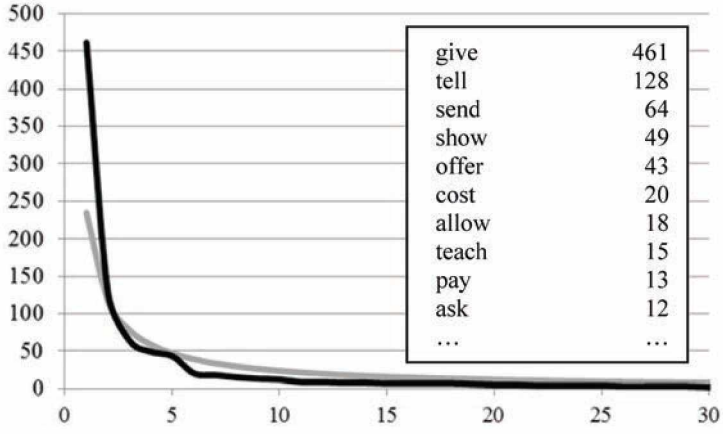


Abb. 11: Verben in der englischen Ditransitivkonstruktion (Rangdarstellung); 30 Typen, 931 Token (Daten aus Stefanowitsch/Gries 2003, S. 229)

(2) a. NP^{nom} V $NP^{akk/dat}$ PP^{an}
b. das gefällt ihr / fasziniert sie / stört sie / ... an ihm

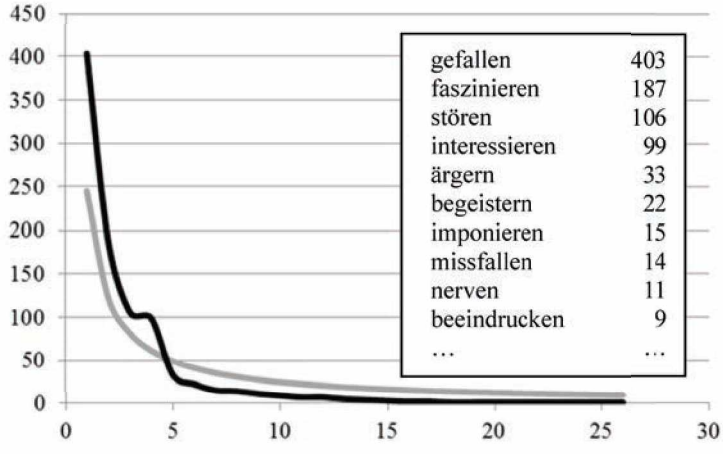


Abb. 12: Verben in der deutschen Split-Stimulus-Konstruktion mit *an* (Rangdarstellung); 26 Typen, 938 Token (Daten aus Engelberg ersch. demn.)

(3) a. NP^{nom} V $Rede^{dir/indir}$
b. sie sagt / erklärt / meint / ..., daß ...
c. sie sagt / erklärt / meint / ...: „...“

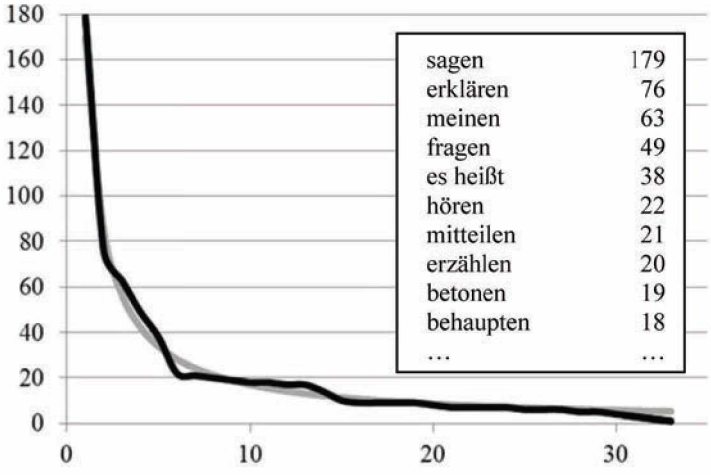


Abb. 13: Verben in direkter oder indirekter Redeeinleitungskonstruktion (Rangdarstellung); 33 Typen, 691 Token (Daten aus Jäger 1968, S. 248 f.)

Die drei Beispiele illustrieren, dass Argumentstrukturmuster oft eine Verteilung aufweisen, die insbesondere bezüglich des rangniedrigsten Verbs (hier *give*, *gefallen* und *sagen*) eine Frequenz deutlich über der Zipf-Verteilung zeigt.

Abweichungen, die in noch stärkerem Maße in die gleiche Richtung weisen, also eine weit über Zipfschen Erwartungen liegende Häufigkeit des rangniedrigsten Elements zeigen, finden sich bei festen Wendungen. Dass die Vorstellung von der Festigkeit von „festen" Wendungen einer erheblichen Korrektur bedarf, hat vor allem Steyer in vielen Studien nachgewiesen (vgl. v.a. Steyer 2013). So zeigt etwa eine Korpusstudie, dass die usuelle Wortverbindung (*Dein*) *Wort in Gottes Ohr* allein an der Position „Gottes" 60 weitere Füllungen aufweist, die alle jeweils nur einmal (56 Füllungen) oder zweimal (4 Füllungen) belegt sind, z.B.: *dein Wort in Allahs/des Politikers/Henkels/des Wählers/Toyotas/... Ohr*. Demgegenüber stehen 73 Belege mit der kanonischen Füllung *Gottes*.[13] Knapp die Hälfte der 137 Korpusbelege zeigen also Varianz allein an dieser einen Position der Wendung. Die Verteilung der Füllwörter entfernt sich allerdings erwartungsgemäß extrem von einer Zipfschen Verteilung (siehe Abb. 14).

13 Wortverbindungen online. Plattform des Projekts Usuelle Wortverbindungen. Internet: wvonline.ids-mannheim.de/wvfelder/koerper/ohr (Stand: 13.8.2013).

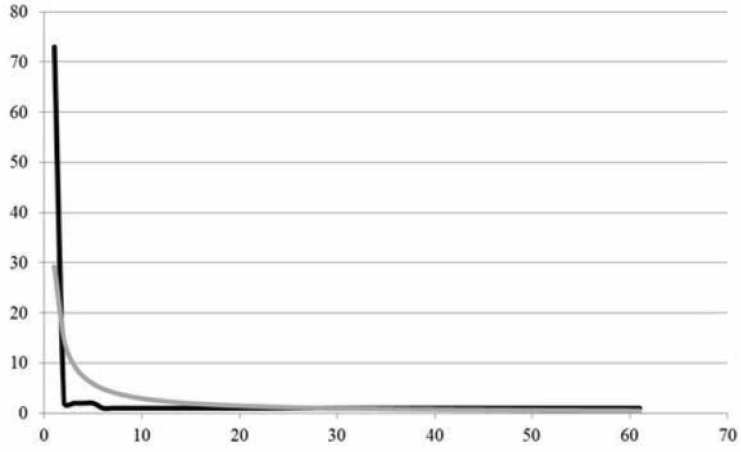


Abb. 14: Verteilung der Slotfüller in (*Dein*) *Wort in __ Ohr*

Die in diesem Abschnitt skizzierten Phänomene zeigen, dass auch funktional zusammenhängende Teilwortschätze und in paradigmatischen Beziehungen stehende Mengen von Slotfüllern in sprachlichen Mustern Verteilungen aufweisen, die dadurch gekennzeichnet sind, dass einige wenige Wörter sehr häufig und viele andere sehr selten auftreten. Es wurde aber auch deutlich, dass die Verteilungen dabei in für die jeweiligen Phänomene charakteristischer Weise von einer Zipfschen Verteilung abweichen.

4 Selbstorganisation und Dynamik

Veränderungen in der Sprache finden Ihren Niederschlag darin, dass sprachliche Formen, Strukturen und Bedeutungen zu einem bestimmten Zeitpunkt in sprachlichen Äußerungen innerhalb bestimmter Kontexte auftreten, in denen sie vor oder nach diesem Zeitpunkt nicht oder nicht in demselben Maße aufgetreten sind. Sprachlicher Wandel hat also mit quantitativen Verteilungen von sprachlichen Einheiten – gegebenenfalls relativ zu bestimmten Kontexten – zu tun. In diesem Abschnitt soll auf den Zusammenhang zwischen Zipfschen Verteilungen und dynamischen Prozessen eingegangen werden.

Die Interpretation von zipfnahen Wortschatzverteilungen muss berücksichtigen, dass sich sehr ähnliche Verteilungsmuster auch in vielen anderen Bereichen finden. Der italienische Ökonom Vilfredo Pareto hat sie bereits bei der Verteilung von Vermögen im England des 19. Jahrhunderts beobachtet (Pareto 1897; siehe

auch Scarrott 1974, S. 403). Statt von Zipfschen Verteilungen wird insofern oft auch von Pareto-Verteilungen gesprochen. Darüber hinaus entspricht die Verteilung von Wörtern relativ zu ihrer Gebrauchshäufigkeit auch der Verteilung[14]

- von deutschen Großstädten relativ zu ihrer Bevölkerungszahl,
- von Wirtschaftsunternehmen relativ zu ihrer Mitarbeiterzahl,
- von Bestsellern relativ zu ihren Verkaufszahlen,
- von Sprachen relativ zu ihrer Sprecherzahl,
- von Mondkratern relativ zu ihrem Durchmesser,
- von Geschäftstagen an der Börse relativ zu ihrer Volatilität,
- von der Stärke von Erdbeben relativ zu ihrer Auftretenshäufigkeit,
- von Asteroiden relativ zu ihrer Größe,
- von wissenschaftlichen Artikeln relativ zu ihrer Zitierhäufigkeit,
- von Familiennamen relativ zur Häufigkeit ihrer Namensträger,
- von Webseiten relativ zu den Klicks, die sie erhalten,
- von Sonneneruptionen relativ zu ihrer Strahlungsintensität
- u.v.m.

Ob sich eine universelle Erklärung für Zipfsche Verteilungen findet, die die verschiedenen Natur-, Gesellschafts- und Kognitionsphänomene erfasst, ist umstritten (siehe dazu Piantadosi 2014, S. 24 ff.). Die Spezifik von Wortschatzverteilungen im Besonderen ist mit verschiedenen pragmatischen, semantischen und kognitiven Erklärungen verknüpft worden.[15] Zipf (1949, S. 20 f.) selber suchte die Erklärung in der Spannung zwischen Sprecherökonomie und Hörerökonomie. Der Sprecher, der ja bei der Konzeption der Aussage Wörter aus dem Lexikon auswählen muss, kann seine Anstrengung umso mehr reduzieren, je kleiner das Lexikon ist. Der Hörer dagegen minimiert seine Verstehensanstrengungen, wenn die Aussage des Sprechers semantisch differenziert ist, was wiederum ein semantisch vielfältiges, umfangreiches Lexikon verlangt.[16]

Plausibel scheint es in diesem Zusammenhang auch, den Spannungsbogen der Zipfschen Kurve einerseits mit dem geringeren Sprachverarbeitungsaufwand zu verknüpfen, der mit hoher Gebrauchshäufigkeit üblicherweise einhergeht, und andererseits mit dem Bedürfnis nach kreativem Sprachgebrauch, wie er sich in lexikalischen Neuerungen und dem Auftreten von Wörtern in ungebräuchlichen Kontexten ausdrückt. Diese Annahme ist auch mit Beobachtungen aus Experi-

14 Vgl. etwa Scarrott (1974, S. 402); Li (2002, S. 16 ff.); Newman (2005, S. 327 ff.); Shirky (2012, S. 272).

15 Verschiedene Erklärungsansätze werden in Piantadosi (2014) dargestellt.

16 Weiterentwickelt wurde dieser Ansatz etwa von Ferrer i Cancho (2005).

menten zu sprachlicher Kreativität kompatibel, in denen Probanden in verbalen Assoziationstests ein unterschiedliches Reaktionsverhalten auf Stimuluswörter zeigen: Weniger kreative Probanden haben starke Assoziationen (operationalisiert als besonders kurze Reaktionszeiten) zu einigen wenigen Wörtern, produzieren aber relativ kurze Wortlisten, während kreative Probanden weniger starke Assoziationen zeigen, dafür aber längere Wortlisten produzieren (Bousfield/Sedgewick/Cohen 1954; Mednick 1962, S. 223; Wallach/Kogan 1965, S. 15 ff.). Setzen wir diese Ergebnisse in Beziehung zu den Zipfschen Verteilungen, so äußert Kreativität sich eben vor allem in der Verlängerung des Schweifs der Zipfverteilung.

Semantische Ansätze wiederum betonen den Zusammenhang von Zipfschen Verteilungen mit dem Netz semantischer Relationen, das sich im Lexikon manifestiert. So versucht Manin (2008) die Zipfsche Verteilung aus dem Prinzip der Synonymievermeidung und der hierarchischen Struktur hyperonymischer Bäume abzuleiten. Darüber hinaus wurden Zipfsche Verteilungen auch mit strukturellen und funktionalen Eigenschaften des Gehirns in Verbindung gebracht, etwa mit der Periodiziät neuronaler Prozesse (Lebedev 1983 nach Tuldava 1998, S. 43).

Typischerweise finden sich Zipf-Pareto-Verteilungen in Systemen, die durch Selbstorganisiertheit,[17] eine hohe Interaktion ihrer Elemente, konstanten Wandel und internen Wettbewerb gekennzeichnet sind:

> The law applies where there is a free competition mechanism: towns competing with one another for people or words competing with one another for social use and storage in a limited human memory. [...] There is good reason to suppose that such an explanation must be founded upon the fact that towns and natural languages are „living" in the sense that they constantly change and interact as a result of many free human decisions. (Scarrott 1974, S. 402)

Wandelprozesse im Lexikon sind somit im Zusammenhang mit Zipfschen Verteilungen zu betrachten. Auch wenn es durchaus Arbeiten zum lexikalischen Wandel auf der Basis Zipfscher Verteilungen gibt (siehe etwa Altmann et al. 2002, S. 134 ff.), sind viele der in den obigen Abschnitten angesprochenen Phänomene bisher mangels hinreichend großer historischer Korpora kaum auf quantitative Aspekte lexikalischen Wandels hin untersucht worden. Beispielhaft seien hier drei Phänomene vorgestellt, deren Untersuchung mit der besseren Verfügbarkeit geeigneter Korpora möglich wird: Argumentstrukturen, feste Wendungen und Quasi-Synonyme.

---

17 Zu einigen Überlegungen zur Selbstorganisation der Sprache vgl. etwa Altmann (1981) und Köhler (1986).

Schaut man sich die Verteilung von Redeeinleitungsverben in Heinrich Wittenwilers satirischem Lehrgedicht „Der Ring" an (Wießner (Hg.) 1931), so sehen wir die für Argumentstrukturmuster typische über den Zipfschen Erwartungen liegende Frequenz des häufigsten Verbs. Verglichen mit der Verteilung in gegenwartssprachlichen Texten (Abb. 13) ist es in dem Text aus der Zeit um 1400 allerdings ein anderes Verb, das die Verteilung dominiert, nämlich *sprechen* an Stelle von *sagen* (Abb. 15). Wie sich die quantitative Dynamik gestaltet, die solch radikale Änderung in zipfnahen Verteilungen hervorbringt, ist eine noch zu untersuchende Frage.

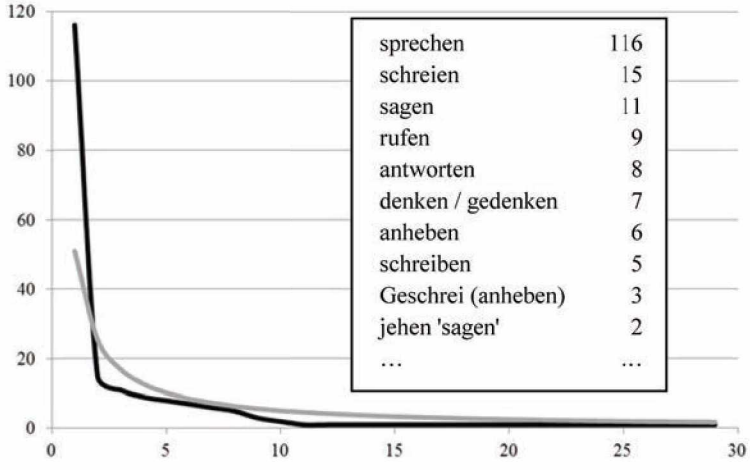


Abb. 15: Verteilung der Verben in der Direkte-Rede-Konstruktion in „Heinrich Wittenwiler: Der Ring", Stichprobe von 202 Token mit 28 Typen

Ein zweites Beispiel für erst im Ansatz erforschte Veränderungen in quantitativen Verteilungen betrifft die Entstehung von Varianten fester Wendungen. Zu dem 1999 populär gewordenen Slogan *„Wir können alles. Außer Hochdeutsch"* haben sich in der Folgezeit eine Vielzahl von Varianten gebildet, entweder in Form einer veränderten syntaktischen Konstruktion oder durch lexikalische Ersetzungen:[18]

(4) a. Ein Mann aus Leonberg vertrat die Auffassung, es sei doch nicht gescheit, wenn einer von sich behaupte, **er könne alles außer Hochdeutsch**. (Frankfurter Allgemeine, 13.11.1999)

18 Vgl. zur Genese und Verwendung dieses Slogans auch König (2013).

b. Für das Bundesland Sachsen hatten Scholz & Friends nämlich vor einiger Zeit mit dem Satz geworben: „**Wessis können alles, außer Sächsisch.**“ (Mannheimer Morgen, 10.12.1999)

c .Baden-Württemberg ist in diesem Bereich – ich spreche jetzt von Integrationspolitik – nicht Spitze, was es ja sonst immer sehr gerne ist. **Wir können alles außer Integration!** (Protokoll der Sitzung des Landtags von Baden-Württemberg, 3.2.2000)

Eine Korpusrecherche ergibt, dass von den 507 in DeReKo ermittelten Belegen 50,49% in der ursprünglichen Form des Slogans auftreten und 49,51% Varianten darstellen. Interessanterweise zeigt die Zeitverlaufsanalyse, dass die Variantenbildung kein spät einsetzender Prozess ist, sondern fast unmittelbar mit dem Auftauchen des Slogans beginnt (siehe Abb. 16). Über die Zeit hinweg bleibt das quantitative Verhältnis zwischen kanonischer und variierter Form in einer Weise stabil, die ein Abhängigkeitsverhältnis nahelegt. Die Varianten benötigen die kanonische Form zur Interpretation ihres sprachspielerischen Impetus, und die kanonische Form scheint von ihrem Potenzial zu kreativer Variantenbildung zu profitieren.

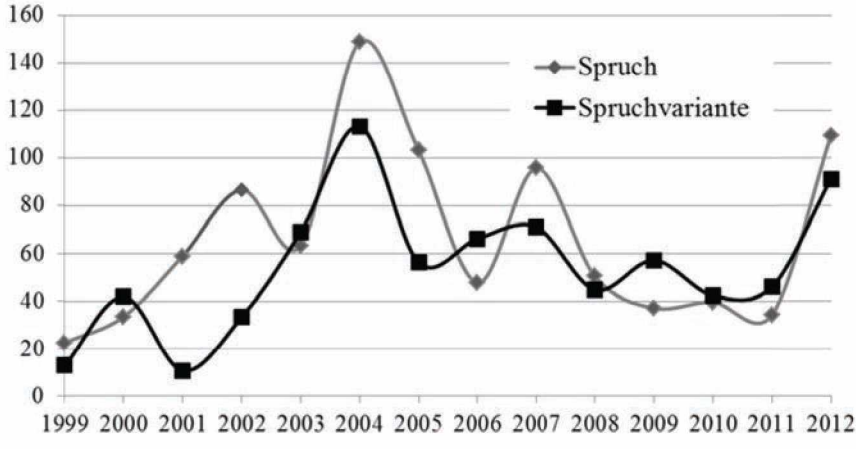


Abb. 16: Zeitverlaufsgrafik zur Entwicklung von Varianten der Wendung *Wir können alles. Außer Hochdeutsch*

Die Dynamik in den lexikalischen Verteilungen findet auch dort ihren Ausdruck, wo Wörter durch ihre enge semantische Verwandtschaft in Wettbewerb miteinander treten und sich in Folge dieses Wettbewerbs ihre Frequenzränge in der Wortschatzverteilung verschieben. So ist etwa aus den Zeitverlaufsgrafiken des IDS-

Neologismenwörterbuchs zu ersehen, dass Anfang der 1990er Jahre drei Wörter zur Bezeichnung eines mobilen Kommunikationsgeräts in Konkurrenz standen: *Handtelefon*, *Mobiltelefon* und *Handy* (siehe Abb. 17).

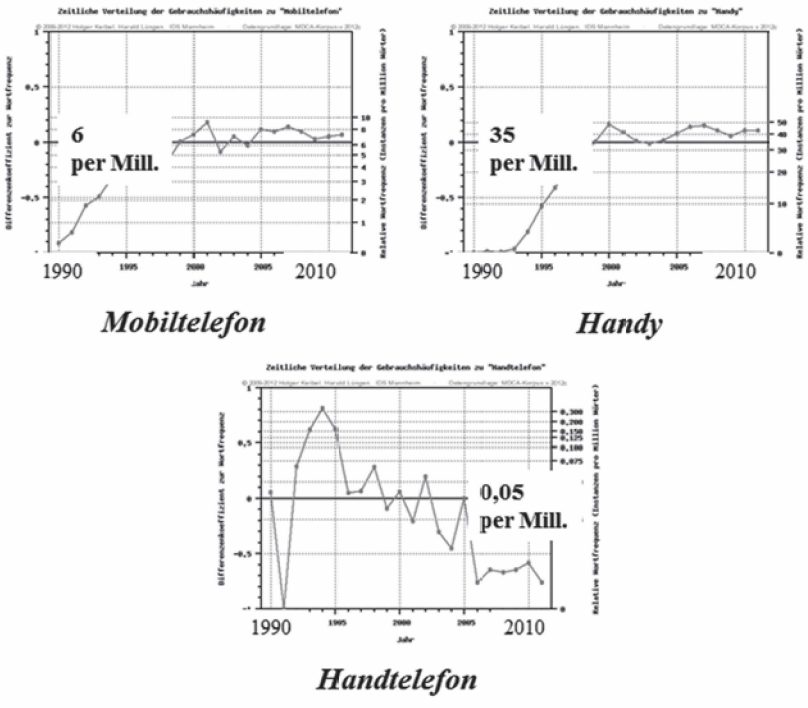


Abb. 17: Zeitverlaufsgrafiken zu *Mobiltelefon*, *Handy*, *Handtelefon* aus dem IDS-Neologismenwörterbuch (bearbeitet von Doris Steffens, Doris al-Wadi). Internet: www.owid.de/wb/neo (Stand: 4.1.2014)

Die Frequenzverläufe zeigen dabei zwei auch in anderen lexikalischen Wettbewerbsverläufen zu beobachtende Muster. Erstens sieht man, dass eine eher formellere Bezeichnung (*Mobiltelefon*) am Anfang der Entwicklung steht, kurz darauf gefolgt von einer informelleren (*Handy*). Beide zeigen eine ähnliche Etablierungskurve mit einer stabilen Situation um 2000, die informelle Bezeichnung allerdings mit höherer absoluter Frequenz. Zweitens treten manchmal Wörter in die Konkurrenz ein, die schon frühzeitig aus dem Wettbewerb wieder ausscheiden (*Handtelefon*). Ähnliche Prozesse lassen sich mit Hilfe des Google-Ngram-Viewers[19]

19 Google-Ngram-Viewer. Internet: https://books.google.com/ngrams (Stand: 11.3.2014).

auch für andere lexikalische Konkurrenzen beobachten. So finden sich vergleichbare Verlaufskurven der Verdrängung eines Quasi-Synonyms auch bei dem Paar *Flugplatz/Aerodrom*, und solche einer sich stabilisierenden stilistischen Varianz formell/informell bei Wortpaaren wie *Automobil/Auto* oder *Fernsehgerät/Fernseher*. Die Daten deuten also auf eine diachrone Musterhaftigkeit verschiedener Typen lexikalischer Wettbewerbe innerhalb semantischer Felder.

Die obigen Beispiele zeigen, dass die sich in den zipfnahen Verteilungen manifestierende Selbstorganisiertheit des sprachlichen Systems mit einer Vielzahl von Phänomenen lexikalischer Dynamik verknüpft ist. Solche Phänomene auf der Basis wachsender diachroner Korpora zu untersuchen, wird eine wichtige Aufgabe der zukünftigen Lexikonforschung sein.

5 Sprachdynamik und Lexikografie

Auch wenn es ein Gemeinplatz ist, dass Sprache ein dynamisches System vorstellt, das permanentem Wandel unterliegt, spiegelt sich diese Überzeugung in synchronen Sprachdokumentationen, seien es Grammatiken oder Wörterbücher, kaum wider. Natürlich gibt es Wörterbücher, die den lexikalischen Wandel post hoc dokumentieren, aber das dynamische Potenzial, das der Wortschatz aus synchroner Perspektive aufweist, drückt sich in Wörterbüchern des Gegenwartsdeutschen kaum aus. Dynamische Aspekte zeigen sich hier bestenfalls in Form diachroner Markierungen (z.B. „veraltet“ oder „obsolet“) oder in der Lemmatisierung von Wortbildungselementen, die zu einem gewissen Grad das Potenzial zu derivationsbasierten Neubildungen ausdrücken.[20]

Natürlich gibt es auch spezielle Wörterbücher, die Aspekten der Dynamik oder der quantitativen Verteilung gewidmet sind, aber diese verknüpfen nicht die in dem vorliegenden Artikel diskutierten Zusammenhänge zwischen Verteilung und Dynamik und dienen insofern auch nicht als Wörterbücher, die den synchronen Wortschatz in seiner Dynamik dokumentieren: Historische Wörterbücher sind einzelwortbezogen und berücksichtigen üblicherweise keine quantitativen Aspekte; Frequenzwörterbücher beschränken sich demgegenüber meist ausschließlich

20 Indirekt spiegelt sich zudem die Frequenz eines Wortes in der Länge der Wörterbuchartikel und dem Grad an Polysemie eines Wortes wider: hochfrequenten, polysemen Wörtern werden im Wörterbuch im Allgemeinen mehr Platz eingeräumt; siehe etwa Prün (2005) zum Verhältnis zwischen der Häufigkeit eines Wortes und der Anzahl seiner Bedeutungen.

auf Frequenzinformationen zu Einzelwörtern; Neologismenwörterbücher sind zwar meist frequenzbasiert, fokussieren aber ebenfalls post hoc auf das Einzelwort und bieten wenig Informationen zu generellen Neuerungstendenzen im Gesamtwortschatz.[21]

In einem anderen Artikel (Engelberg 2014, S. 251) habe ich für ein Wörterbuch der Sprachdynamik plädiert, das Sprache in ihren lexikalischen Aspekten und im Übergang zur Grammatik als ein dynamisches System darstellt. Lexikalische Dokumentationen der Gegenwartssprache sollten sich demnach an den Begriffen der Musterbildung, der Variation, der quantitativen Verteilungen, der Exemplarbasiertheit, der Dynamik und der Vernetzung orientieren.

Der vorliegende Artikel soll einen weiteren Schritt darstellen zur Entwicklung des Konzepts eines dynamischen Lexikons als Gegenstand lexikalischer Sprachdokumentationen: Das dynamische Lexikon repräsentiert das synchrone Potenzial zu lexikalischem Wandel, zu lexikalischer Variation und zu lexikalischem Wachstum. Das dynamische Lexikon ist:

- exemplarbasiert: Jede Äußerung (in einem Korpus) zählt; jede Äußerung formt das dynamische Potenzial des Lexikons.
- sprachmusterbasiert: Neben Ein- und Mehrwortausdrücken sind auch (semi-) abstrakte sprachliche Muster Teil des Lexikons (z.B. Wortbildungsmuster, Wortverbindungsmuster und Argumentstrukturmuster).
- verteilungsmusterbasiert: Die Veränderungspotenziale lexikalischer Einheiten sind an quantitative Verteilungsmuster geknüpft, die bestimmte Phänomene lexikalischen Wandels charakterisieren.

Insbesondere der letzte Aspekt stand in dem vorliegenden Artikel im Mittelpunkt. Dabei charakterisierte sich Folgendes heraus:

- Die Größe des Lexembestands gegenwärtiger Korpora liegt um ein bis zwei Größenordnungen über dem Bestand der in großen Wörterbüchern lemmatisierten Lexeme.
- Der Modalwert in Frequenzverteilungen von Lexemen in Korpora ist 1; d.h., das typische Lexem ist ein Hapax Legomenon.
- Quantitative Verteilungsmuster charakterisieren bestimmte Typen von Phänomenen (Verben in Argumentstrukturen, Funktionswörter, Kompositabildung, feste Wendungen).

---

21 Im IDS-Neologismenwörterbuch (Doris Steffens/Doris al-Wadi, www.owid.de/wb/neo) werden über die Zeitverlaufsgrafiken Frequenzinformationen bereitgestellt, und die erweiterte Suche bietet zudem bestimmte Möglichkeiten, die Entwicklungen in bestimmten Wortschatzbereichen zu verfolgen.

– Im Zeitverlauf zeigen sich Typen von quantitativen Veränderungen, die vermutlich charakteristisch für bestimmte Sprachwandelphänomene sind (Herausbildung stilistischer Varianz bei Quasi-Synonymen, Änderungen in Verbpräferenzen von Argumentstrukturen, Entstehung von Varianz bei festen Wendungen).

Stimmt man der These zu, dass diese Eigenschaften elementare Charakteristika eines synchronen Wortschatzsystems sind, stellen sich für die internetbasierte Lexikografie der Sprachdynamik ganz neue Aufgaben: Insbesondere sind vonnöten (i) eine Fokussierung auf funktional zusammenhängende Teilwortschätze, die über das semantische Verweissystem der einzelwortorientierten Lexikografie hinausgeht, (ii) die Entwicklung von Visualisierungen, die es erlauben, die quantitativen Zusammenhänge im Gesamtwortschatz und in Teilwortschätzen erfahrbar zu machen, (iii) die Entwicklung von Präsentationsformen, die es ermöglichen, die für den Wortschatz so charakteristischen seltenen Lexeme in die Dokumentation miteinzubinden und (iv) natürlich eine Einbettung der Beschreibung lexikografischer Einheiten in ein Format, das dessen synchrones dynamisches Potenzial zum Ausdruck bringt.

Literatur

Altmann, Gabriel (1981): Zur Funktionalanalyse in der Linguistik. In: Esser, Jürgen/Hübler, Axel (Hg.): Forms and functions. (= Tübinger Beiträge zur Linguistik 149). Tübingen, S. 25–32.

Altmann, Gabriel et al. (2002): Einführung in die quantitative Lexikologie. Göttingen.

Baayen, R. Harald (2001): Word frequency distributions. Dordrecht.

Bousfield, Weston A./Sedgewick, C. H. W./Cohen, B. H. (1954): Certain temporal characteristics of the recall of verbal associates. In: The American Journal of Psychology 67, S. 111–118.

Duden (1999): Großes Wörterbuch der deutschen Sprache. 10 Bde. 3., völlig neu bearb. u. erw. Aufl. Hrsg. v. Wissenschaftlichen Rat der Dudenredaktion. Mannheim u.a.

Engelberg, Stefan (2014): Gegenwart und Zukunft der Abteilung Lexik am IDS: Plädoyer für eine Lexikographie der Sprachdynamik. In: Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim, S. 243–253.

Engelberg, Stefan (2015): The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In: Boas, Hans/Ziem, Alexander (Hg.): Constructional approaches to argument structure in German. Boston/Berlin.

Engelberg: Stefan (ersch.demn.): Gespaltene Stimuli bei Psych-Verben: Kombinatorische Mustersuchen in Korpora zur Ermittlung von Argumentstrukturverteilungen. In: Engelberg, Stefan et al. (Hg.): Argumentstruktur – Valenz – Konstruktionen. Tübingen.

Evert, Stefan/Baroni, Marco (2005): Testing the extrapolation quality of word frequency models. In: Proceedings of „Corpus Linguistics", Birmingham, July 14-17, 2005. Internet: www.corpus.bham.ac.uk/PCLC/ (Stand: 23.5.2014).

Ferrer i Cancho, Ramon (2005): The variation of Zipf's Law in human language. In: The European Physical Journal B – Condensed Matter and Complex Systems 44, S. 249–257.
Jäger, Siegfried (1968): Die Einleitungen indirekter Reden in der Zeitungssprache und in anderen Texten der deutschen Gegenwartssprache. In: Muttersprache 78, S. 236–249.
Klein, Wolfgang (2013): Von Reichtum und Armut des deutschen Wortschatzes. In: Deutsche Akademie für Sprache und Dichtung/Union der Deutschen Akademien der Wissenschaften (Hg.): Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Berlin/Boston, S. 15–55.
Köhler, Reinhard (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. (= Quantitative Linguistics 1). Bochum.
Köhler, Reinhard/Altmann, Gabriel/Piotrowski, Rajmund G. (Hg.) (2005): Quantitative Linguistik. Ein internationales Handbuch. (= Handbücher zur Sprach- und Kommunikationswissenschaft 27). Berlin/New York.
König, Werner (2013): „Wir können alles. Außer Hochdeutsch." Genialer Werbespruch oder Eigentor des deutschen Südens? Zum Diskriminierungspotential dieses Slogans. In: Sprachreport 2013/4, S. 5–14.
Lebedev, A. N. (1983): Zakonomernosti povtorenija slov v reči. In: Psichologičeskij žurnal 5, S. 11–22.
Li, Wentian (2002): Zipf's Law everywhere. In: Glottometrics 5, S. 14–21.
Mandelbrot, Benoit (1961): On the theory of word frequencies and on related Markovian models of discourse. In: Jakobson, Roman (Hg.): Structure of language and its mathematical suspects. Proceedings of the Twelfth Symposium in Applied Mathematics, held in New York City, April 14-15, 1960. Providence, RI, S. 190–219.
Manin, Dmitrii (2008): Zipf's Law and avoidance of excessive synonymy. In: Cognitive Science 32, S. 1075–1098.
Mednick, Sarnoff A. (1962): The associative basis of the creative process. In: Psychological Review 69, S. 220–232.
Naranan, Sundaresan/Balasubrahmanyan, Vriddhachalam K. (2005): Power laws in statistical linguistics and related systems. In: Köhler/Altmann/Piotrowski (Hg.), S. 716–738.
Newman, Mark E. J. (2005): Power laws, Pareto distributions and Zipf's Law. In: Contemporary Physics 46, S. 323–351.
Pareto, Vilfredo (1897): Cours d'économie politique. Teil 2. Lausanne u.a.
Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Paderborn.
Piantadosi, Steven T. (2014): Zipf's Word Frequency Law in natural language: A critical review and future directions. University of Rochester. Internet: http://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf (Stand: 27.4.2014).
Prün, Claudia (2005): Das Werk von G.K. Zipf. In: Köhler/Altmann/Piotrowski (Hg.), S. 142–152.
Rosengren, Inger (1972): Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung. Bd. 1. (= Lunder Germanistische Forschungen 41). Lund.
Rosengren, Inger (1977): Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung. Bd. 2. (= Lunder Germanistische Forschungen 43). Lund.
Scarrott, G. S. (1974): Will Zipf join Gauss? In: New Scientist, 16 May 1974, S. 402–404.
Shirky, Clay (2012): Das Pareto-Prinzip. In: Brockman, John (Hg.): Was macht uns schlauer? Frankfurt a.M., S. 270–273.
Stadler, Heike (erscheint. demn.): Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora. (= OPAL – Online publizierte Arbeiten zur Linguistik 5/2014). Mannheim.

Stefanowitsch, Anatol/Gries, Stefan T. (2003): Collostructions: Investigating the interaction of words and constructions. In: International Journal of Corpus Linguistics 8, S. 209–243.
Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. (= Studien zur Deutschen Sprache 65). Tübingen.
Tuldava, Juhan (1998): Probleme und Methoden der quantitativ-systemischen Lexikologie. (= Quantitative Linguistics 59). Trier.
Wahrig, Gerhard (1980–1984): Brockhaus-Wahrig. Deutsches Wörterbuch. 6 Bde. Wiesbaden.
Wallach, Michael A./Kogan, Nathan (1965): Modes of thinking in young children. A study of the creativity-intelligence distinction. New York u.a.
Wießner, Edmund (Hg.) (1931): Heinrich Wittenwilers Ring. Nach der Meininger Handschrift. (= Deutsche Literatur. Reihe Realistik des Spätmittelalters 3). Leipzig.
Zipf, George Kingsley (1949): Human behavior and the principle of least effort. Cambridge, MA.