Zipfsches Gesetz

Zusammenhang zwischen Rang, Häufigkeit und Wortlänge

In seinem Buch „Human Behaviour and the Principle of Least Effort" (Zipf X) skizziert George Kingsly Zipf eine Theorie menschlicher Sprache und Gesellschaft, die sich am Effizienzbegriff der Ökonomie orientiert und ein Kosten-Nutzen-Optimierendes „Prinzip des geringsten Aufwands" in den Mittelpunkt stellt. Ausgangspunkt seiner Überlegung ist die Frage, wie das Vokabular einer Sprache strukturiert sein muss, damit das Verhältnis zwischen der Anzahl von Wörtern und der Anzahl von Bedeutungen, die jedes Wort hat, aus Sicht des Sprechers und aus Sicht des Hörers einer Sprache optimal ist. Seinem Ansatz nach wäre für einen Sprecher der Kodierungsaufwand dann am geringsten, wenn er so wenig Wörter wie möglich verwenden würde, auch wenn diese Wörter viele Bedeutungen haben, während für einen Hörer der Aufwand für die Dekodierung dann am geringsten ist, wenn er nur mit eindeutigen Wörtern konfrontiert würde, auch wenn dies wiederum sehr viele wären. Zipf spricht von diesen beiden Aspekten als zwei im Widerstreit stehenden Kräften, der *Force of Unification* und der *Force of Diversification*. Seine zentrale Annahme ist nun, dass das Vokabular der gesprochenen Sprache eine (optimale) Balance zwischen genau diesen zwei Kräften darstellt (Zipf S. 22). Die Kraft der Unifikation bewirkt dabei eine Reduzierung der Anzahl der Wörter verbunden mit einer Erhöhung der Häufigkeit ihrer Verwendung, wogegen die Kraft der Diversifikation in umgekehrter Richtung eine Erhöhung der Anzahl der Wörter verbunden mit einer Verminderung der Häufigkeit ihrer Verwendung bewirkt. Er folgert daraus, dass die Anzahl von Wörtern und die Häufigkeit ihrer Verwendung die zentralen Parameter für die Beschreibung der Vokabularstruktur einer Sprache sind: "number and frequency will be the parameters of vocabulary balance" (Zipf S.23).

Seine Überlegung illustriert Zipf am Beispiel des Vokabulars von James Joyce *Ulysses* (unter Verwendung des sogenannten *Hanly Index*). Dabei sind alle Wortformen, die im Text vorkommen, ihrer Häufigkeit nach in eine geordnete Liste geschrieben:

Hanly Index von James Joyce Ulysses (aus Zipf 1949, S. 24), Auszug

I	II	III	IV
Rank r	Frequenz f	Produkt aus I und II r * f = k	Theoretische Länge von "Ulysses" K * 10
10	2.653	26.530	265.500
20	1.211	26.220	262.200
30	926	27.780	277.800
50	556	27.800	278.800
100	265	26.500	265.000
500	50	25.500	250.000
1.000	26	26.000	260.000
2.000	12	24.000	240.000
5.000	5	25.000	250.000
10.000	2	20.000	200.000
20.000	1	20.000	200.000
29.899	1	29.899	298.990

Wie man erkennt, ist das Produkt aus dem Rang einer Wortform (innerhalb der häufigkeitssortierten Liste) mit ihrer Häufigkeit in etwa konstant. Dieser Zusammenhang wird allgemein als das Zipfsche Gesetz bezeichnet.

Zipfsches Gesetz: $r * f \sim k$ (mit Rang *r*, Frequenz *f* und korpusspezifischer Konstante *k*)

Wird das Zipfsche Gesetz in einem Funktionsgraph mit Rang *r* als x-Achse und Frequenz *f* als y-Achse dargestellt, erhält man eine Hyperbel. Anders ausgedrückt: Die Häufigkeit des Auftretens einer Wortform im Text ist umgekehrt proportional zu ihrem Rang.

Wählt man für die x- und y-Achse eine doppelt logarithmische Darstellung, also *log(r)* und *log(f)*, dann erhält man mit $\log(f) = -\log(r) + \log(k)$ eine Gerade mit negativer Steigung –1, wie es Abbildung 1 am Beispiel eines deutschen Korpus mit 222 Millionen tokens verdeutlicht. (Für die Berechnung der Geraden sind dabei die ersten 10 häufigsten types ebenso wie der long tail von types mit Frequenz 1 nicht berücksichtigt worden).

Werden reale Textdaten verwendet, dann entspricht dieses ideale Verhältnis von Rang und Häufigkeit jedoch nicht dem tatsächlichen Funktionsverlauf. Der Zusammenhang zwischen Rang und Häufigkeit wird für Wortformen mit sehr kleinem oder sehr großem Rang nur unzureichend durch die einfache Formel $n \sim 1/r$ wiedergegeben. Im Diagramm mit logarithmisch skalierten Achsen weichen diese Wortformen stärker von der vorausgesagten Geraden ab.

Eine bessere Beschreibung liefert folgende Formel nach Benoît B. Mandelbrot (vgl. Mandelbrot **Fehler! Verweisquelle konnte nicht gefunden werden.**):

$$(r + c_1)^{1+c2} * f \sim k$$

Die beiden Konstanten c_1 und c_2 dienen hier als Parameter. Sie ermöglichen eine Anpassung an die konkreten Daten. Setzt man $c_1 = c_2 = 0$, ergibt sich die ursprüngliche Formel von Zipf.

Wie Abbildung 2 verdeutlicht, liefern die Parameter $c_1 = 6{,}5$, $c_2 = 0{,}22$ und $k = 80\,000\,000$ (als die am besten approximierten Werte) eine bessere Prognose für die Daten des Projekts Deutscher Wortschatz. Hierbei hat c_1 großen Einfluss auf die Krümmung im Bereich der niederen Ränge, während c_2 die Anpassung im Bereich der hohen Ränge vornimmt.

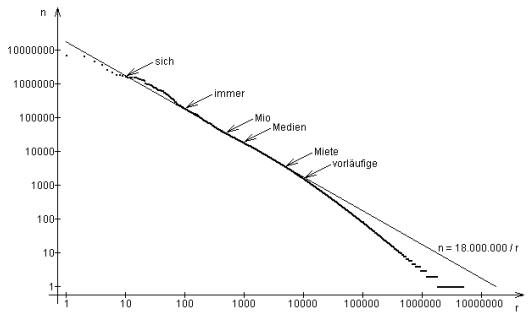


Abbildung 1

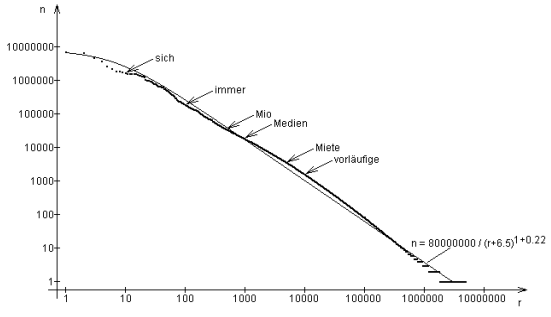


Abbildung 2

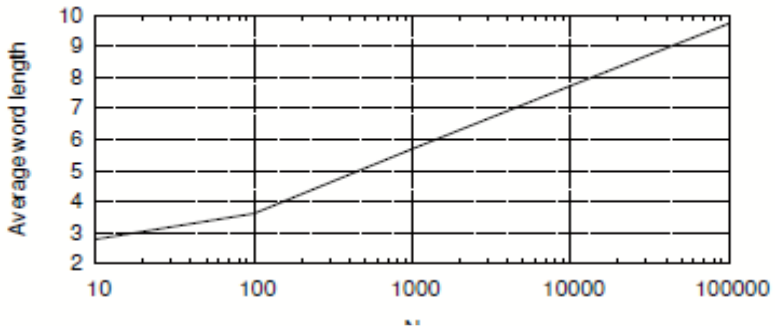
Das „Prinzip des geringsten Aufwands" findet sich aber auch in der Kodierung der Wortformen wieder. So sind zum Beispiel die am häufigsten gebrauchten Wortformen meist sehr kurze Funktionswörter und seltener verwendete Wortformen sind länger. Fürs Deutsche gilt dabei, dass eine Wortform um nahezu zwei Buchstaben länger wird, wenn sich der Rang um den Faktor 10 erhöht (Eckart et.al. **Fehler! Verweisquelle konnte nicht gefunden werden.**).

10 häufigste Worte des Projekts Deutscher Wortschatz (Quelle: Deutscher Wortschatz, deu_newscrawl-public_2019_100K, 1.498.643 tokens, 145.971 types)

Rang r	Wortform	Häufigkeit n	r*n
1	der	44.066	44.066

2	die	42.150	84.300
3	und	32.366	97.098
4	in	25.799	103.196
5	den	17.083	85.415
6	das	13.627	81.762
7	mit	13.227	92.589
8	ist	13.147	105.176
9	zu	13.083	117.747
10	von	12.816	128.160

Korrelation der durchschnittlichen Wortlänge mit dem Rang (in logarithmischer Skalierung), Steigung der Geraden 1,99, vgl. Eckart et. al. 2011, S. 18)



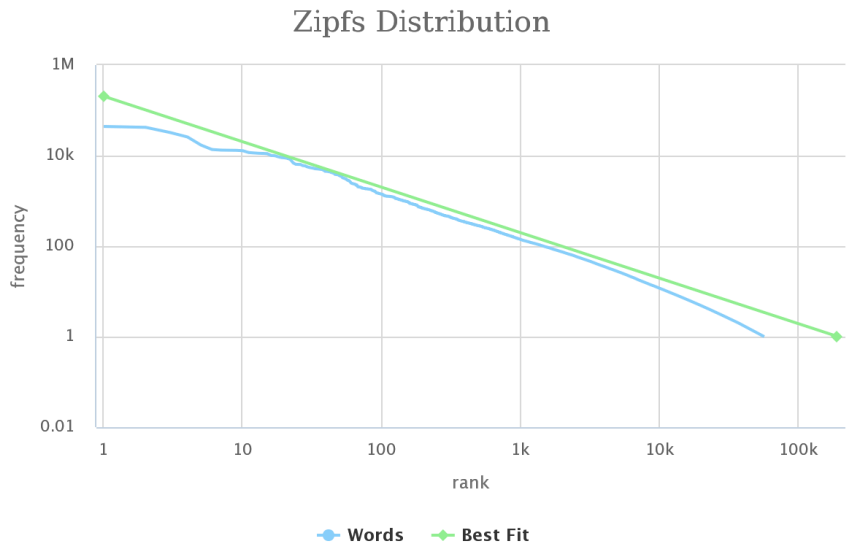
Die Zipf-Verteilung von Wortformen in einem Text können Sie online mit der ASV-Toolbox berechnen (vgl. Anhang 5, ASV-Toolbox **Fehler! Verweisquelle konnte nicht gefunden werden.**). Gehen Sie dazu auf die Seite

https://toolbox.wortschatz.uni-leipzig.de/toolbox/#!tools/languagestatistics

und geben Sie einen beliebigen Text ein oder wählen Sie ein Beispiel aus den für dieses Buch bereitgestellten Referenzkorpora.

Berechnung der Zipf-Verteilung mit der ASV-Toolbox mit Daten deu_newscrawl-public_2019_100K)

Rang gegen Häufigkeit, Projekt Deutscher Wortschatz, best fit: -1.0781279625283025x+204698



Sprachabhängige Konstante c, Vorhersagen und Anwendungen

Die textabhängige Konstante ***k*** ist abhängig von der Korpusgröße. Durch die Normalisierung dieses Parameters auf die Anzahl der in einem Korpus insgesamt vorkommenden tokens erhalten wir die sprachabhängige Konstante ***c***, die für alle Korpora einer Sprache gelten sollte.

Die sprachabhängige Konstante ***c*** berücksichtigt anstelle der absoluten Häufigkeit eines Wortes seine relative Häufigkeit und wird bestimmt über:

$$r * f/N = k/N \sim c$$

Aus den oben genannten Daten des Projekts Deutscher Wortschatz mit N = 222 538 789 tokens und k = 18 000 000 ergibt sich so beispielsweise für das Deutsche eine Konstante von

$$c = 18.000.000/222.539.789 \sim 0.08$$

Die sprachabhängige Konstante c sowie die Steigung der Ausgleichsgeraden a ist unterschiedlich für verschiedene Sprachen wie die nachfolgende Tabelle verdeutlicht.

Sprachspezifische Konstanten c und Steigung a auf Basis der Daten web_public_2019_10K

Sprache	c	a
Deutsch	0.0898	-1.040
Englisch UK	0.0935	-1.034
Finnisch	0.1058	-0.812
Arabisch	0.1140	-0.989
Chinesisch	0.1238	-0.979

Empirische Untersuchungen großer Textkorpora zeigen, dass sich die Steigung der Ausgleichsgraden für eine Sprache eines bestimmten Genres auch bei zunehmender Korpusgröße nur wenig verändert. Für größere Korpora eines bestimmten Genres einer Sprache erhalten wir daher in der doppelt logarithmischen Darstellung der Zipf-Verteilung parallele Geraden, die sich mit zunehmendem Umfang nur in dem Parameter **b = log (k)** unterscheiden.

Sind für ein Textkorpus die sprachspezifische Konstante c und die Gesamtzahl der tokens gegeben, dann können aus dem Zipfschen Gesetz eine Reihe von Voraussagen zum Verhältnis von Rang und Frequenz von types in diesem Korpus abgeleitet werden.

Nehmen wir an, dass für ein Textkorpus bereits eine frequenzsortierte Liste erzeugt worden ist. Am Anfang der Liste stehen die types mit der höchsten Frequenz und dem kleinsten Rang, diejenigen types, die nur 1 mal im Text vorkommen, stehen am Ende der Liste. Das letzte Element der Liste bekommt den höchsten Rang. Jeder type kommt zwar n mal im Textkorpus vor, für ein bestimmtes n gibt es aber möglicherweise mehrere types, die genau so oft im Korpus vorkommen.

Betrachten wir als erstes den höchsten Rang von types, die genau f mal in der frequenzsortierten Liste eines Korpus vorkommen. Wir schreiben dafür r_f und wenden das Zipfsche Gesetz an. Es gilt demnach $r_f * f/N = c$. Der höchste Rang von n types, die genau f mal vorkommen ist also

$$r_f = c * N/f$$

Höchster Rang von Wortformen, die genau 50 mal in einem Text vorkommen

Wenn ein deutschsprachiger Text einen Umfang von N=150.000 Wortformen hat, dann befindet sich die letzte Wortform, die genau 50 mal im Text vorkommt, in der häufigkeitssortierten Wortformenliste ungefähr an Position $r_{50} = 0{,}08 * (150.000/50) = 240$.

Für den type mit Rang 1, also der häufigsten Wortform, gilt:

$$1_f = c * N$$

Der höchste Rang derjenigen types, die nur 1 mal vorkommen, markiert die Gesamtzahl von types im Korpus, also den Umfang v des Vokabulars, und ist gegeben durch:

$$r_1 = v = c * N$$

Umfang des Vokabulars einer Textkollektion: $r_1 = v = c * N$

Wir können den Rang eines types verwenden, um abzuschätzen, wie oft ein type in einem Korpus vorkommt (cf. Salton 1989). Wir schreiben für die Anzahl n von tokens von einem type n_f. Diese Anzahl entspricht genau der Differenz zwischen dem höchsten Rang derjenigen types, die in der frequenzsortierten Liste genau f mal vorkommen, und dem höchsten Rang derjenigen types, die genau f+1 mal vorkommen. Daher gilt:

$$n_f = r_f - r_{f+1} = c * N/f + c * N/f+1 = c * N/ f(f+1)$$

Für types mit Frequenz 1 macht das Zipfsche Gesetz die Vorhersage, dass ihr Anteil die Hälfte aller types eines Korpus ausmacht:

$$n_1 = r_1 - r_2 = c * N/1 * (1+1) = c * N/2$$

Der Umfang des Vokabulars v wächst mit der Textgröße N. Diese Beziehung lässt sich abschätzen durch:

$$t = kN^{\beta}$$

Für das Projekt Deutscher Wortschatz gilt k = 20 und β = 0,648. Damit lässt sich voraussagen, dass bei Erweiterung der Textmenge etwa jeder 70. type zum ersten Mal gesehen wird.

Validität

Das Zipfsche Gesetz bildet die Grundlage für zahlreiche sprachstatistische Untersuchungen wie beispielsweise den Korpusvergleich mit Hilfe der type-token-ratio, die Kookkurrenzanalyse oder Topic Modelle. Umso wichtiger ist es deshalb, sich einige Einschränkungen seiner Gültigkeit vor Augen zu führen.

- c ist nicht nur unterschiedlich zwischen einzelnen Sprachen, sondern auch stark abhängig von der Korpusgröße allerdings ist c in etwa konstant, wenn wir mit c = k/log N anstatt c = k/N rechnen (neues Bild), hilft auch bei type-token-ratio
- dennoch: in den einzelnen Frequenzbereichen (Rängen, logarithmisch) finden wir unterschiedlich starke Abweichung (absolut) der Vorhersagen von den realen Werten, diese Muster sind sehr charakteristisch für einzelne Sprachen und sie verstärken sich mit zunehmender Korpusgröße