Sprachstatistik: Das Zipf'sche Gesetz

Korpus-Workshop 29.09.2005

Thema „Korpus-Statistik“

Thomas Wittig

Universität Leipzig

Institut für Informatik

wittig@uni-leipzig.de

Principle of Least Effort

George K. Zipf: Für natürliche Sprache gilt das „Prinzip der geringsten Anstrengung“.

z.B. sind die am häufigsten gebrauchten Wörter einer Sprache meist sehr kurze, inhaltsleere Funktionswörter.

Beispiel: 10 häufigste Wortformen aus Projekt „Deutscher Wortschatz“

Wortform	Häufigkeit
der	7.377.879
die	7.036.092
und	4.813.169
in	3.768.565
den	2.717.150
von	2.250.642
zu	1.992.268
das	1.983.589
mit	1.878.243
sich	1.680.106

Daten in den Beispielen

kommen aus Projekt „Deutscher Wortschatz“

Konzept: Sammlung von Texten zur Schaffung einer großen Datenbasis für weitere Untersuchungen

Material: vorwiegend (online verfügbare) Archive von Zeitungen, aber auch News-Ticker, Lexika (z.B. Encarta, Rocklexikon 1999) oder Fachtexte (z.B. SAP-Korpus)

Stand im März 2001

- 24.788.212 Sätze
- mit 222.538.789 Wortformen (tokens)
- darunter 5.122.776 verschiedene Wortformen (types)

1 Die Formel

- Wortformen eines Textes absteigend nach Häufigkeit ordnen

→ Rang r einer Wortform in der Liste multipliziert mit seiner Häufigkeit n ist in etwa konstant.

$r \times n \approx k$ (mit textabhängiger Konstante k)

bzw. ausgedrückt durch

$n \sim 1/r$

häufigkeitssortierte Liste

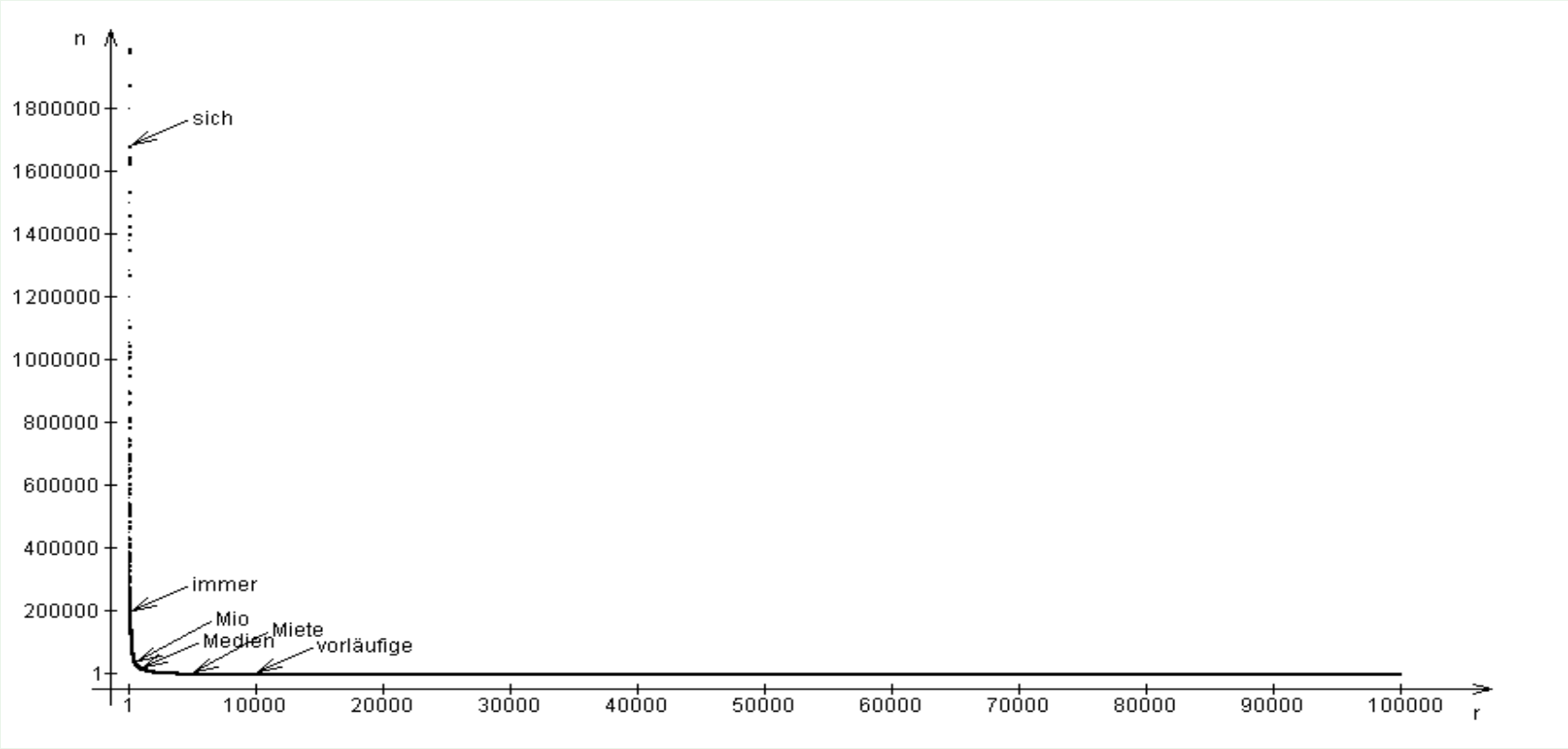
Rang	Wortform	Anzahl
1	der	7377897
2	die	7036092
3	und	4813169
4	in	3768565
5	den	2717150
6	von	2250642
7	zu	1992268
8	das	1983589
9	mit	1878243
10	sich	1680106
11	des	1646885
12	auf	1640124
13	für	1638774
14	ist	1633510
15	im	1626923
16	%##%	1539957
17	dem	1464909
	...	

Rang	Wortform	Anzahl
673	spielt	27455
674	Westen	**27427**
675	Interesse	**27427**
676	verloren	27413
677	Preis	27353
	...	
15133	Engholm	1001
15134	Dummheit	1001
15135	ond	**1000**
15136	Zweig	**1000**
15137	Rein	**1000**
15138	Oberbürgermeisters	**1000**
15139	Käthe	**1000**
15140	Auswirkung	**1000**
15141	Ausschau	**1000**
15142	zählende	999
15143	vorgehaltener	999
	...	

$r \times n \approx k$

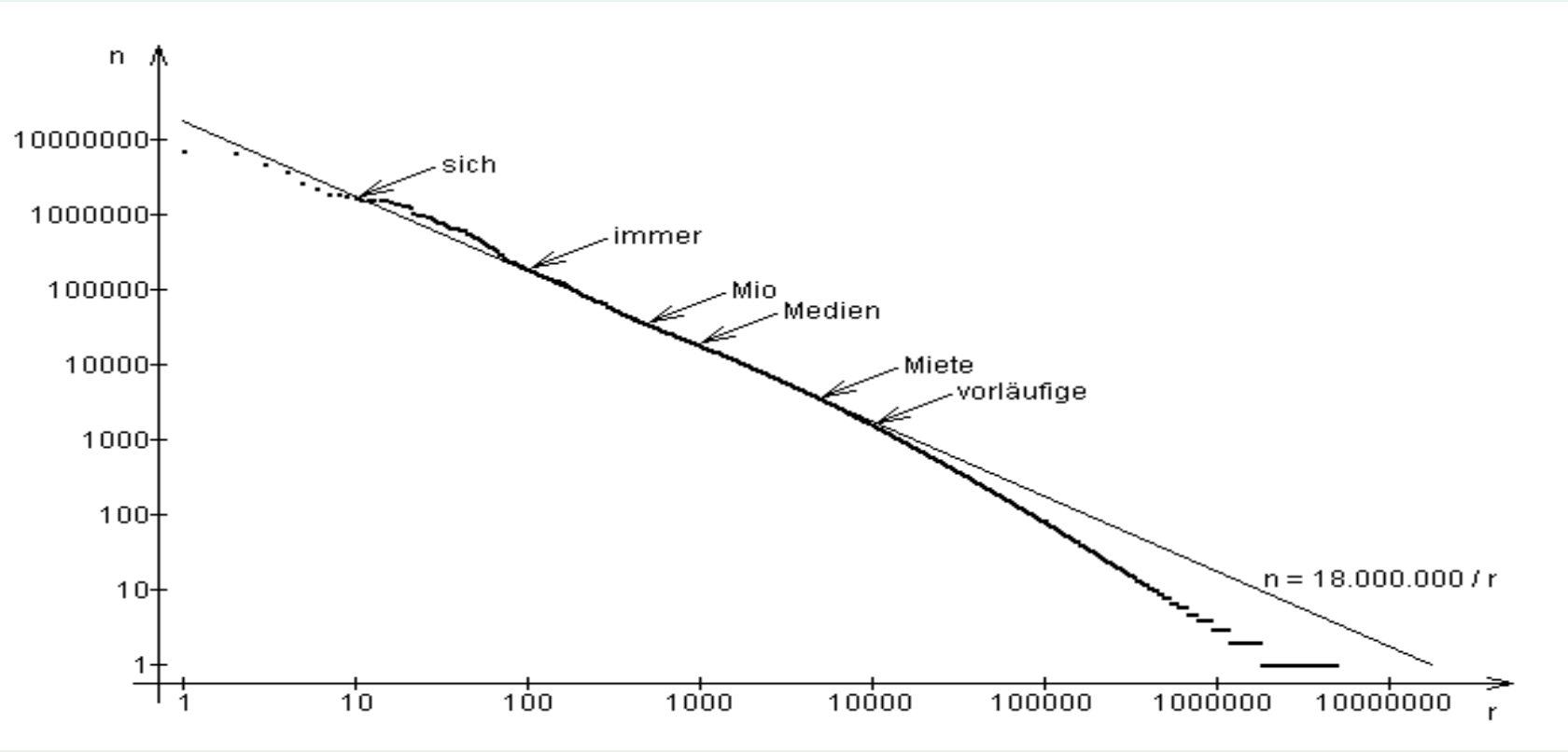
Wortform	Häufigkeit n	Rang r	r × n
sich	1.680.106	10	16.801.060
immer	197.502	100	19.750.200
Mio	36.116	500	18.059.500
Medien	19.041	1.000	19.041.000
Miete	3.755	5.000	18.775.000
vorläufige	1.664	10.000	16.640.000

Graphische Darstellung



Graphische Darstellung

logarithmische Skalierung der Achsen



2 Anwendungen

gegeben: ein konkreter Text

- Abschätzung über Anzahl an Wortformen, die n mal im Text vorkommen (→)

  Welche Textmenge wird benötigt, damit mindestens x verschiedene Wortformen darin vorkommen, die genau n mal auftreten? (←)

- Abschätzung des Umfangs des Vokabulars

- Abschätzung des Zuwachses des Vokabulars, wenn sich Textmenge erhöht

Symbolkonventionen

N Gesamtanzahl aller Wortformen des Textes (tokens)

t Umfang des Vokabulars (types)

n/N relative Häufigkeit einer Wortform, die n mal auftritt

r_n größter Rang derjenigen Wortformen, die genau n mal auftreten

I_n Anzahl der Wortformen, die genau n mal auftreten

→ Es gilt (nach Salton 1989):

$r_n \times n/N = c$ (Konstante c ist textunabhängig, aber sprachabhängig)

Abschätzung für r_n

Beispiel: Ein deutschsprachiger Text bestehe aus 150.000 Wortformen (tokens).

Die Position der letzten Wortform, die 50 mal im Text vorkommt, in der häufigkeitssortierten Wortformenliste lässt sich folgendermaßen abschätzen:

$$r_n \times n/N = c$$

$$r_n = c \times N/n$$

$$r_{50} = 0{,}08 \times 150.000/50 = 240$$

Die Anzahl der Wortformen des Textes, die mindestens 50 mal im Text vertreten sind, lässt sich also auf 240 schätzen.

Abschätzung des Vokabulars

Für t, den Umfang des Vokabulars, gilt:

t ist so groß wie der größte Rang der häufigkeitssortierten Liste. Falls Wörter mit Häufigkeit 1 vorkommen folgt damit:

$$t = r_1 = c \times N/1 = c \times N$$

am Beispiel

$$t = 0{,}08 \times 150.000 = 12.000$$

Abschätzung der Konstante c

Für die sprachabhängige Konstante c gilt:

$c = r_{(n)} \times n/N$

$= k/N$ (nach Zipf'schem Gesetz)

Nach den Daten des Projekts „Deutscher Wortschatz" gilt damit fürs Deutsche:

$c = 18.000.000 / 222.000.000 \approx 0.08$

Abschätzung für I_n (1)

Für I_n, die Anzahl der Wortformen, die genau n mal auftreten, gilt:

$$I_n = r_n - r_{n+1} = c \times N/n - c \times N/(n+1) = cN/(n(n+1))$$

am obigen Beispiel

$$I_{50} = 0{,}08 \times 150.000/(50 \times 51) \approx 5$$

Für I_1 gilt insbesondere:

$$I_1 = cN/(1 \times 2) = t/2$$

→ Die Hälfte des Vokabulars eines Textes tritt wahrscheinlich nur ein einziges mal auf.

Abschätzung für I_n (2)

allgemein:

Anteil der Wortformen, die genau n mal auftreten, am Vokabular eines Textes

$I_n/t = (t/(n(n+1))) / t = 1/(n(n+1))$

am Beispiel

$I_{50}/t = 1/(50 \times 51) \approx 0{,}04\ \%$

Abschätzung des Wachstums von t

Das Wachstum des Vokabulars, wenn sich die Textmenge erhöht, lässt sich abschätzen mit (nach Salton 1989):

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

Für das Projekt „Deutscher Wortschatz“ gilt $k = 20$ und $\beta = 0.648$ (approximierte Werte).

→ Voraussage: Bei Erweiterung der Textmenge wird etwa jedes 70. Wort zum ersten mal gesehen.

Weitere Zusammenhänge

Beziehungen, die für den Großteil der Wortformen eines Textes gelten, wobei einzelne Wortformen zum Teil deutlich abweichen können:

- Bezeichne l die Länge einer Wortform, dann gilt:
  $n \sim 1/l$ *(wird nicht von den Daten bestätigt)*

- Bezeichne m die Anzahl der Bedeutungen einer Wortform, dann gilt (nach Zipf 1945):
  $m \sim 1/\sqrt{r}$

Gesetzmäßigkeit ist nicht auf Verteilung von Wörtern in Texten beschränkt:
Für Ordnung von amerikanischen Städte nach Einwohnerzahl gilt:
Rang $\times$ Einwohnerzahl $\approx$ konstant

3 Verbesserung

n ~ 1/r beschreibt für Wortformen mit sehr kleinem oder sehr großem Rang nur unzureichend den Zusammenhang zwischen Rang und Häufigkeit

bessere Beschreibung liefert nach B. Mandelbrot (Mandelbrot 1953):

$$n \sim 1/(r+c_1)^{1+c_2}$$

bzw.

$$(r+c_1)^{1+c_2} \times n \approx k$$ (mit textabhängiger Konstante k)

Parameter c_1 und c_2 ermöglichen Anpassung an die konkreten Daten.

Parameter c_1 und c_2

Parameter c_1 und c_2 ermöglichen Anpassung an die konkreten Daten.

c_1: Krümmung im Bereich der niederen Ränge

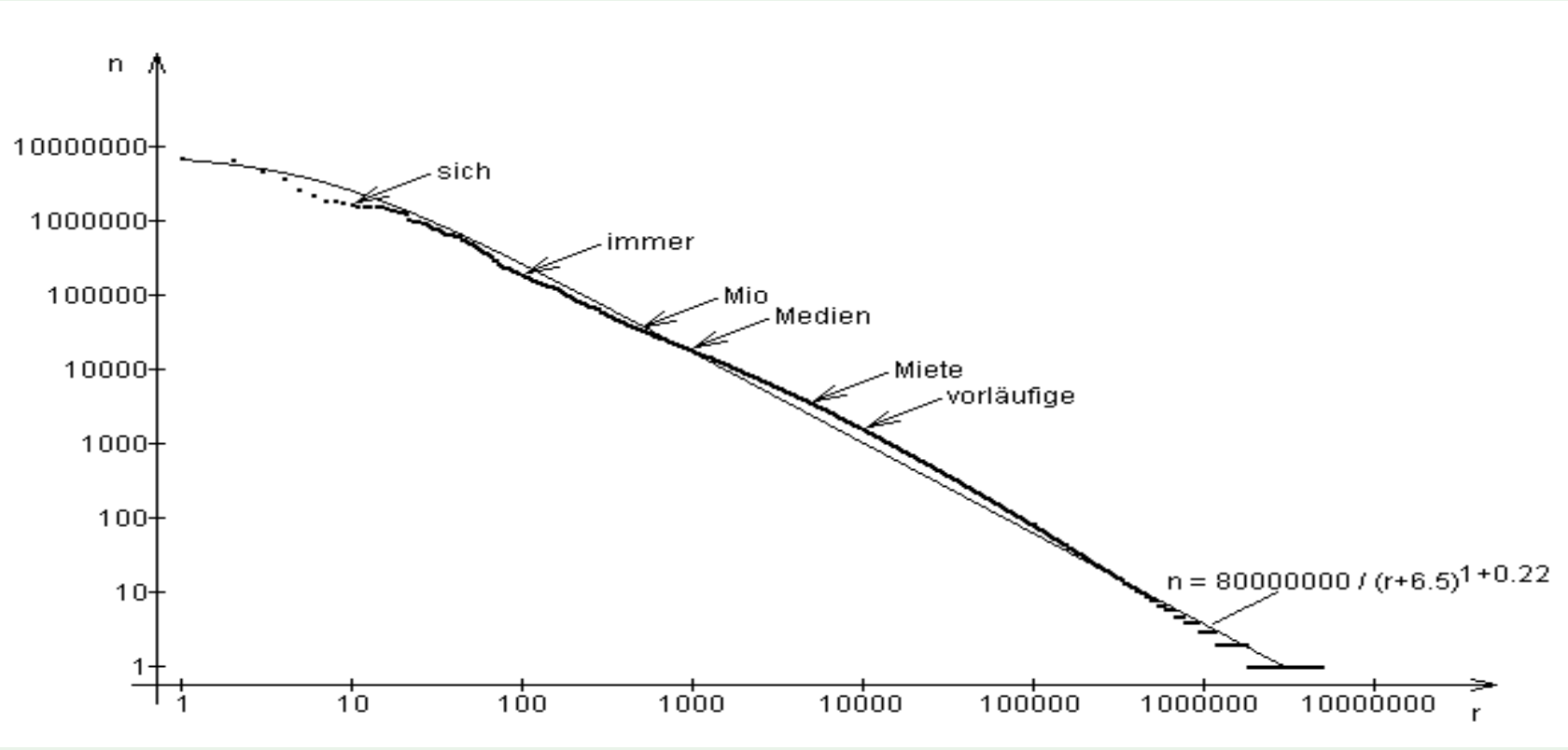
c_2: Anpassung im Bereich der hohen Ränge

$c_1 = c_2 = 0$ ergibt ursprüngliche Formel von Zipf

$$(r+0)^{1+0} \times n \approx k$$

$c_1 = 6{,}5$ und $c_2 = 0{,}22$ (approximierte Werte) liefern bessere Vorhersagen für die Daten des Projekts „Deutscher Wortschatz“.

Graphische Darstellung



Literatur

- Mandelbrot, Benoît B. (1953): An information theory of the statistical structure of language; in: W. Jackson (Ed.), Communication Theory (pp. 503-512). New York: Academic Press.
- Salton, Gerard (1989): Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley.
- Zipf, George K. (1935): The Psycho-Biology of Language. An Introduction to Dynamic Philology. Boston: Houghton-Mifflin.
- Zipf, George K. (1941): The national unity and disunity; The nation as a bio-social organism. Bloomington/Ind.: Princeton Press.
- Zipf, George K. (1945): The meaning – frequency relationship of words, J. Gen. Psycho. 33, 251-266.