Entropie von Sprachen mit lateinischer Schrift

$$H = -\sum_i p_i \log p_i$$

Alexander Guthmann

Information in unserer Sprache

Der Informationsgehalt eines Buchstaben kann über die Differenz der Entropien der (n+1)-Gramme und der n-Gramme definiert werden. [1,2,3] Man kann diese Differenz bedingte Entropie nennen. H ergibt sich als Grenzwert der bedingten Entropien.

$$F_n = -\sum_i p\left(b_i^{n+1}\right) log_2\, p\left(b_i^{n+1}\right) + \sum_i p(b_i^n) \log_2 p(b_i^n)$$

$$H = \lim_{n\to\infty} F_n$$

Wobei:

b_i^n der i-te Block aus n Buchstaben (n-Gramm) ist.
$p(b_i^n)$ die Wahrscheinlichkeit für das i-te n-Gramm ist.

Maximale Entropie M_n von n-Grammen

- Alle Symbole (n-Gramme) kommen vor und sind gleich wahrscheinlich.
- Mit lateinischem Alphabet und Leerzeichen $p_n = \frac{1}{27^n}$

$$M_n = -\log_2 p_n$$

Relative Entropie G_n:

- Maß für maximal mögliche Kompression ohne Informationsverlust mit gegebenen Alphabet
- H_n ist die Entropie der n-Gramme

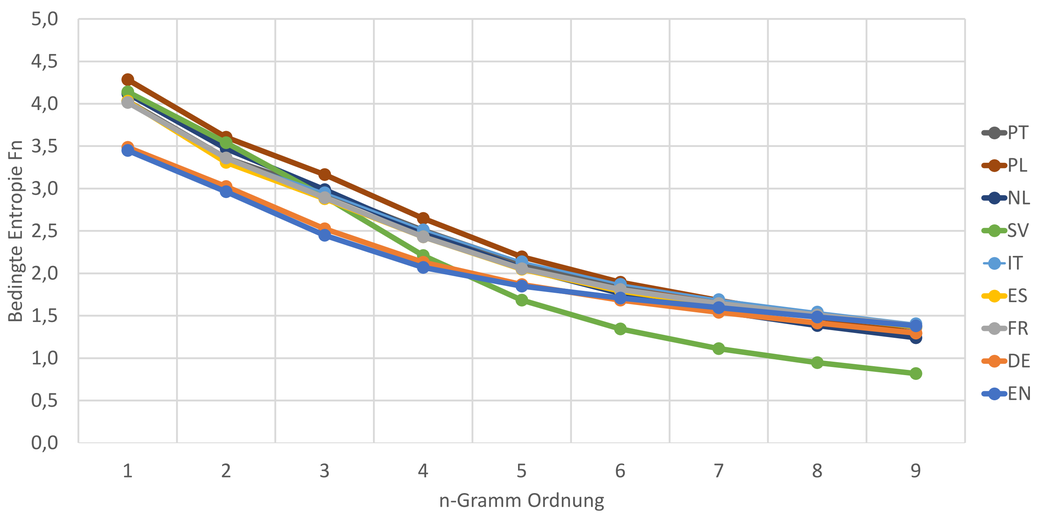
$$G_n = \frac{H_n}{M_n}$$

Redundanz

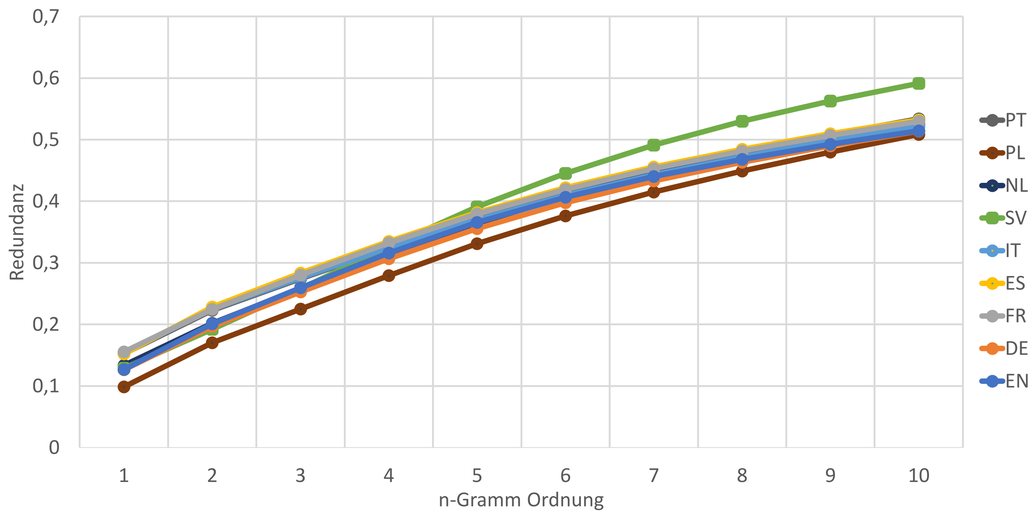
- Maß für Anteil ohne Informationsgehalt

$$R_n = 1 - G_n$$

Berechnete Entropie und Redundanz



➢ Die Information pro Buchstabe liegt bei etwa 1 Bit



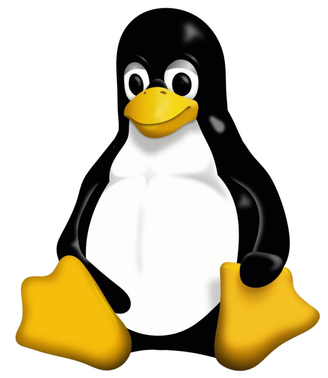
➢ Die Redundanz liegt etwas über 50%

	EN	DE	FR	ES	IT	SV	NL	PL	PT
Größe [GB]	13,0	6,1	5,5	3,8	3,2	2,4	1,8	1,8	1,7
Artikel $* 10^6$	5,8	2,3	2,0	1,5	1,5	3,7	2,0	1,3	1,0

Software

Allgemein:

- ✓ Effizient dank Programmierung in C++
- ✓ Flexible Bedienung mittels Parameterübergabe
- ✓ Platform unabhänig
- ✓ Offener Quellcode
- ✓ Linux Makefile und Windows EXE verfügbar



Programm zum Zählen der n-Gramme:

- ✓ Hashtabelle zum Indizieren der n-Gramme
- ✓ Keine Begrenzung der n-Gramm Größe
- ✓ Zählen von Wörtern möglich
- ✓ Export als Standard CSV-Datei

Programm zum Berechnen der Entropie:

- ✓ Liest CSV-Datei mit n-Gramm Frequenz
- ✓ Keine Beschränkung der n-Gramm Anzahl

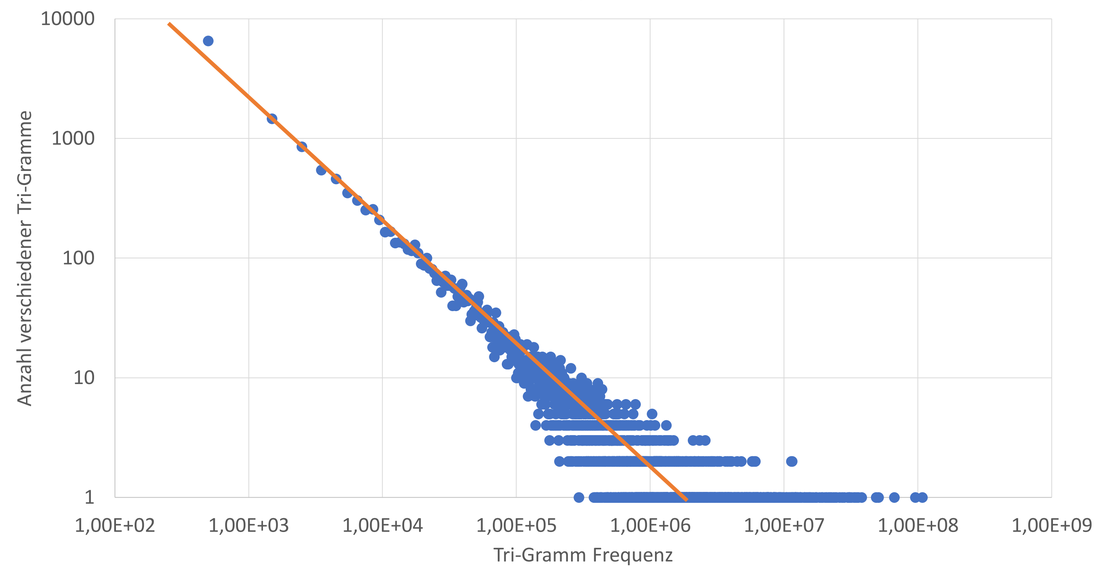
Programm zum Berechnen der Histogramme:

- ✓ Liest CSV-Datei mit n-Gramm Frequenz
- ✓ Beliebige Klassenbreite

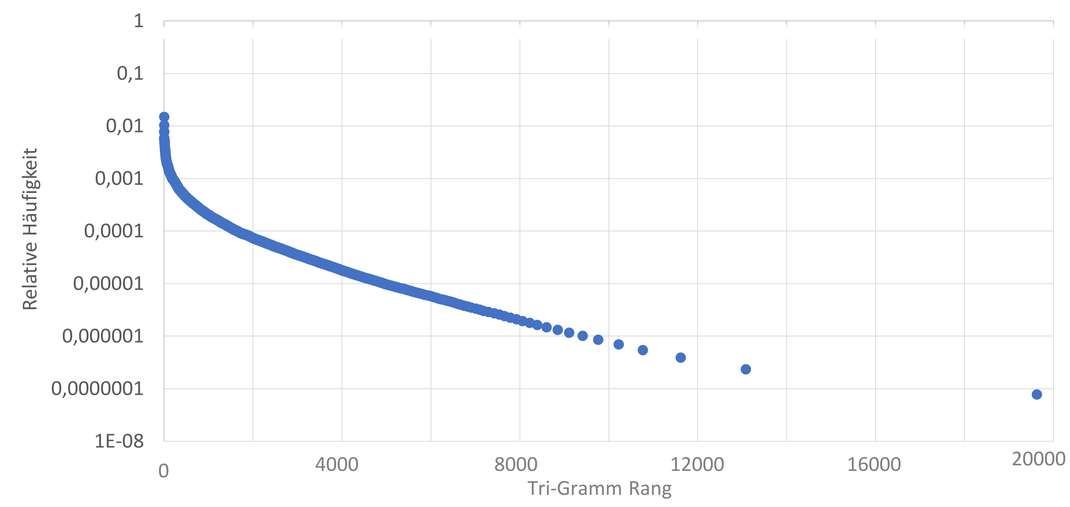
Zipfsches Gesetz und 1/f Rauschen

➢ 1/f Rauschen bzw. das Zipfsche Gesetz sind schon lange bekannt[5]
➢ Bemerkenswert viele Prozesse zeigen das 1/f Verhalten[6]

$$f(r) = \frac{a}{r^\gamma}$$



➢ Tri-Gramm Frequenz der deutschen Sprache
➢ Klassenbreite = 1000
➢ Mehr zum 1/f Verhalten findet sich im Paper



Wikipedia als Datenquelle

- ✓ 295 verschieden Sprachen
- ✓ Qualitativ hochwertiger Text. Keine Probleme wie bei OCR.
- ✓ Liberale und klare Lizenzierung
- ✓ Komplette Datenbank als Dump verfügbar[4]
- ✓ Parser für Dump verfügbar

Alexander Guthmann
TU Kaiserslautern
Email: guthmann@rhrk.uni-kl.de

1. C.E. Shannon, Prediction and Entropy of Printed English, *Bell System Technical Journal* ***30*** *379-423 (1951)*
2. L.B. Levitin, Z. Reingold, Entropy of Natural Languages: Theory and Experiment, *Chaos Solitons & Fractals* ***4****, 709-743 (1994)*
3. C.E. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal* ***27*** *379-423, 623-656 (1948)*
4. https://dumps.wikimedia.org/
5. G. K. Zipf, The Psycho-biology of Language: An Introduction to Dynamic Philology. *Houghton Mifflin Company (1935)*
6. W. H. Press Flicker Noises in Astronomy and Elsewhere *Comments Astrophys.* ***7*** *103-119 (1978)*