Prawo Zipfa
zjawiska (II)

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Temat dzisiejszego odcinka

W jakim stopniu rzeczywista zależność ranga-częstość dla słów odbiega od wyidealizowanej przez Zipfa zależności funkcyjnej?

1. Prawo Zipfa (przypomnienie)
2. Odstępstwa danych od prawa Zipfa (wprowadzenie)
3. Liczba różnych słów w tekście
4. Częstości najczęstszych słów
5. Wzór Mandelbrota
6. Prawo Zipfa dla dużych korpusów
7. Podsumowanie

Lista rangowa

Weźmy sobie pewien tekst (lub korpus tekstów).

Lista rangowa słów to lista słów **posortowanych malejąco** względem **liczby wystąpień** w tekście (korpusie).

Przykład listy rangowej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
1	14767	w
2	12473	i
3	11093	się
...	...	...

Częstość słowa to **liczba wystąpień** słowa w tekście (korpusie).

Ranga słowa to **liczba porządkowa** słowa na liście rangowej:

- najczęstsze słowo ma rangę równą 1,
- drugie co do częstości ma rangę równą 2,
- trzecie co do częstości ma rangę równą 3, ... itd.

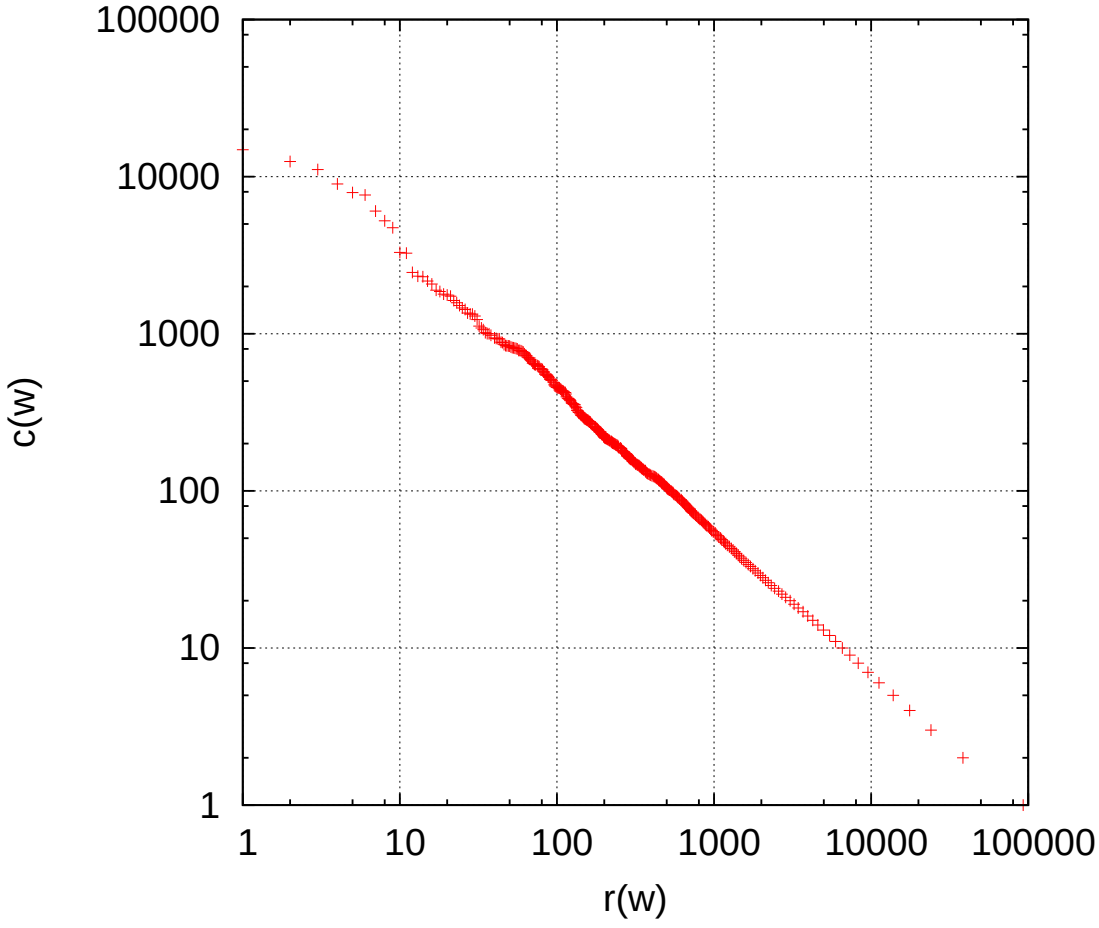
Ciąg dalszy listy rangowej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
...	...	...
4	8750	na
5	7878	nie
6	7605	z
7	6004	do
8	5233	to
9	4675	że
10	3292	jest
11	3264	o
12	2407	jak
13	2320	a
...	...	...

Wykres w skali logarytmicznej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Prawo Zipfa

Jeżeli słowo $\mathbf{w_1}$ ma rangę 10 razy większą niż słowo $\mathbf{w_2}$,
to słowo $\mathbf{w_1}$ ma częstość 10 razy mniejszą niż słowo $\mathbf{w_2}$.

Równoważnie:

Częstość słowa jest odwrotnie proporcjonalna do jego rangi,

$$\mathbf{c(w)} \approx \left\lfloor \frac{\mathbf{A}}{\mathbf{r(w)}} \right\rfloor,$$

gdzie:

$\mathbf{A}$ — liczba różnych słów w tekście,
$\lfloor \mathbf{x} \rfloor$ — część całkowita $\mathbf{x}$, tzn. $\mathbf{x - 1} \leq \lfloor \mathbf{x} \rfloor \leq \mathbf{x}$.

Najsłynniejsze prawo lingwistyki kwantytatywnej

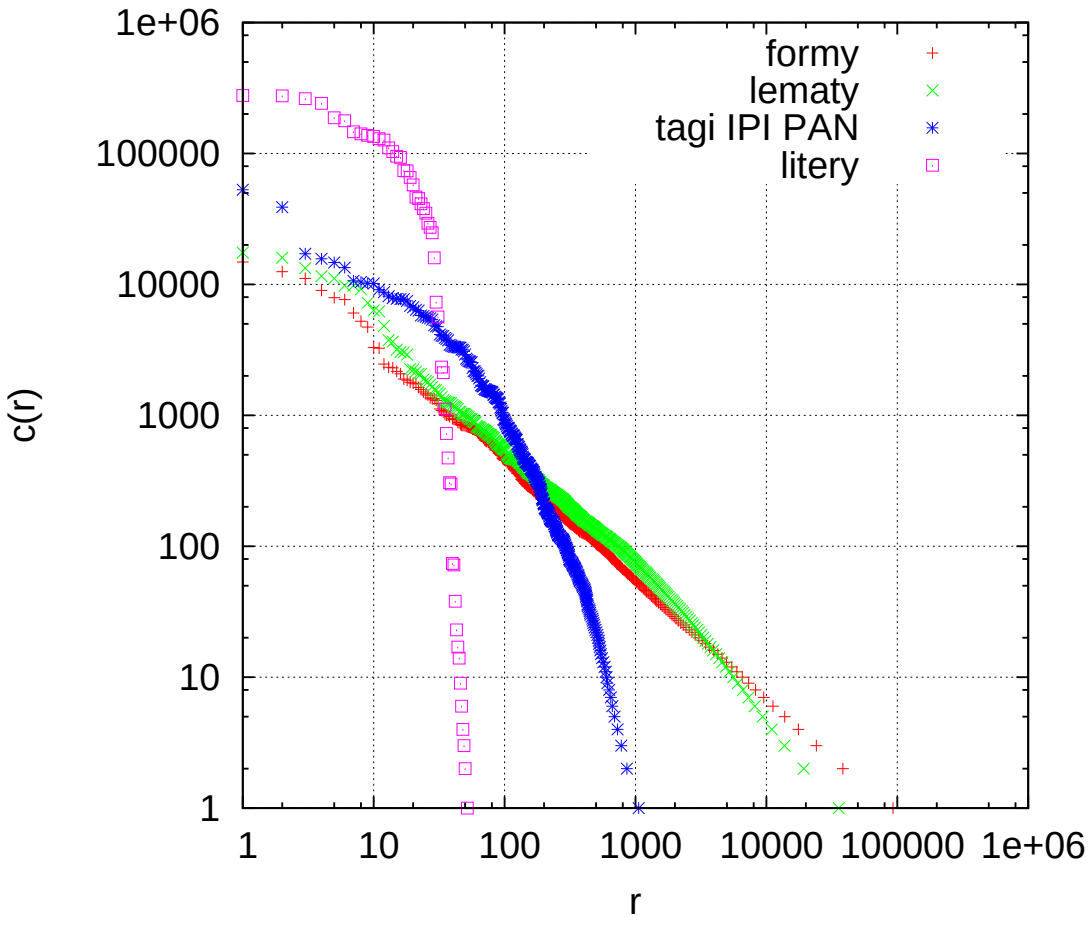
Podobne wykresy dla list rangowych słów obserwuje się dla tekstów pisanych w innych językach.

Nadal brak jest dopracowanego modelu matematycznego, który wyjaśniałby pojawianie się prawa Zipfa.

Wykresy dla listy rangowej liter wyglądają inaczej!

Wykresy ranga-częstość dla różnych obiektów

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Łatwo dostępna informacja o literaturze

interdyscyplinarna bibliografia Wentiana Li

http://www.nslij-genetics.org/wli/zipf/

bibliografia lingwistyki kwantytatywnej A. Pawłowskiego

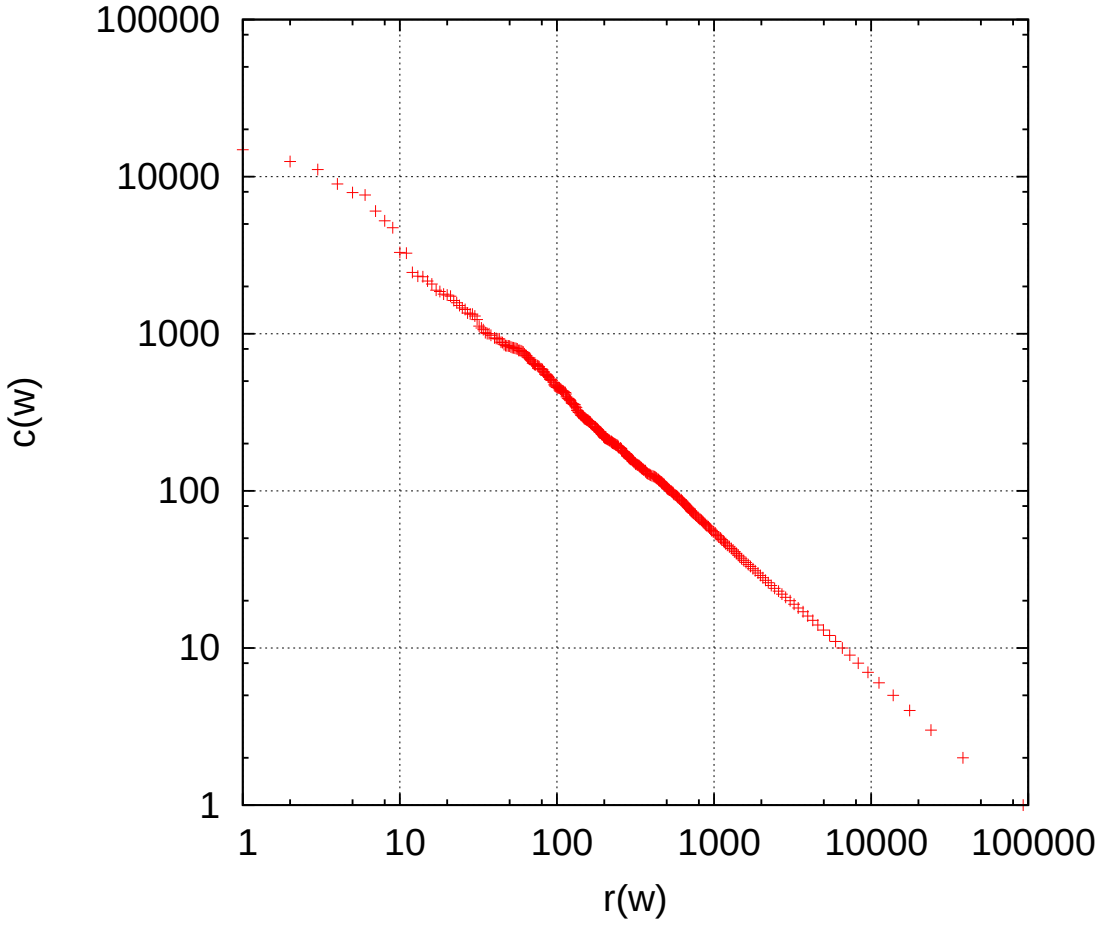
http://www.lingwistyka.uni.wroc.pl/bql/

Temat dzisiejszego odcinka

W jakim stopniu rzeczywista zależność ranga-częstość dla słów odbiega od wyidealizowanej przez Zipfa zależności funkcyjnej?

Wykres w skali logarytmicznej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Jak dobrze ten korpus spełnia prawo Zipfa?

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	**A/r(w)c(w)**	słowo **w**
1	14 767	6,29	w
3	12 497	2,47	się
10	3297	2,81	jest
30	1299	2,38	tego
100	455	2,04	lat
300	157	1,97	mamy
1000	54	1,72	mogły
3000	20	1,55	tworzy
10 000	6	1,55	kursy
30 000	2	1,55	obroną
92 963	1	1,00	aa

Hipoteza robocza

Odstępstwa od prawa Zipfa są systematyczne.

Dwie liczby słów w tekście

A — liczba różnych słów (typów) w tekście
N — liczba wszystkich słów (okazów) w tekście

Obie liczby można wyliczyć mając listę frekwencyjną:

c(r) — częstość słowa o randze **r**

Liczba **A** to największa ranga słowa:

$$\mathbf{c(A) > 0}, \quad \mathbf{c(A + 1) = 0.}$$

Z kolei

$$\mathbf{N = c(1) + c(2) + c(3) + ... + c(A).}$$

Oszacowanie z prawa Zipfa

Prawo Zipfa głosi, że

$$\mathbf{A/r - 1 \leq c(r) \leq A/r.}$$

Liczba wszystkich słów to

$$\mathbf{N = c(1) + c(2) + c(3) + ... + c(A).}$$

Skądinąd wiadomo, że

$$\mathbf{\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{A} \approx \gamma + \ln A \approx 0,577 + 2,30 \cdot \log_{10} A.}$$

Stąd mamy

$$\mathbf{A(\gamma - 1 + \ln A) \ \leq \ N \ \leq \ A(\gamma + \ln A).}$$

Jak dobrze zgadza się to oszacowanie?

$$\mathbf{A(\gamma - 1 + \ln A) \leq N \leq A(\gamma + \ln A)}$$

Korpus Słownika Frekwencyjnego Polszczyzny Współczesnej:

Rzeczywista liczba typów: $\mathbf{A = 92\,963}$

Oszacowanie liczby okazów z prawa Zipfa i liczby typów:

$$\mathbf{A(\gamma + \ln A) \approx 92963(0,577 + 11,439) \approx 1\,117\,000}$$
$$\mathbf{A(\gamma - 1 + \ln A) \approx 1\,117\,000 - 93\,000 \approx 1\,024\,000}$$

Rzeczywista liczba okazów: $\mathbf{N = 555\,073}$

Liczba różnych słów w funkcji długości tekstu

Jeżeli tekst spełnia prawo Zipfa, to

$$\mathbf{N} \leq \mathbf{A}(\gamma + \ln \mathbf{A}).$$

Z drugiej strony, z definicji liczb **A** i **N** mamy

$$\mathbf{A} \leq \mathbf{N}.$$

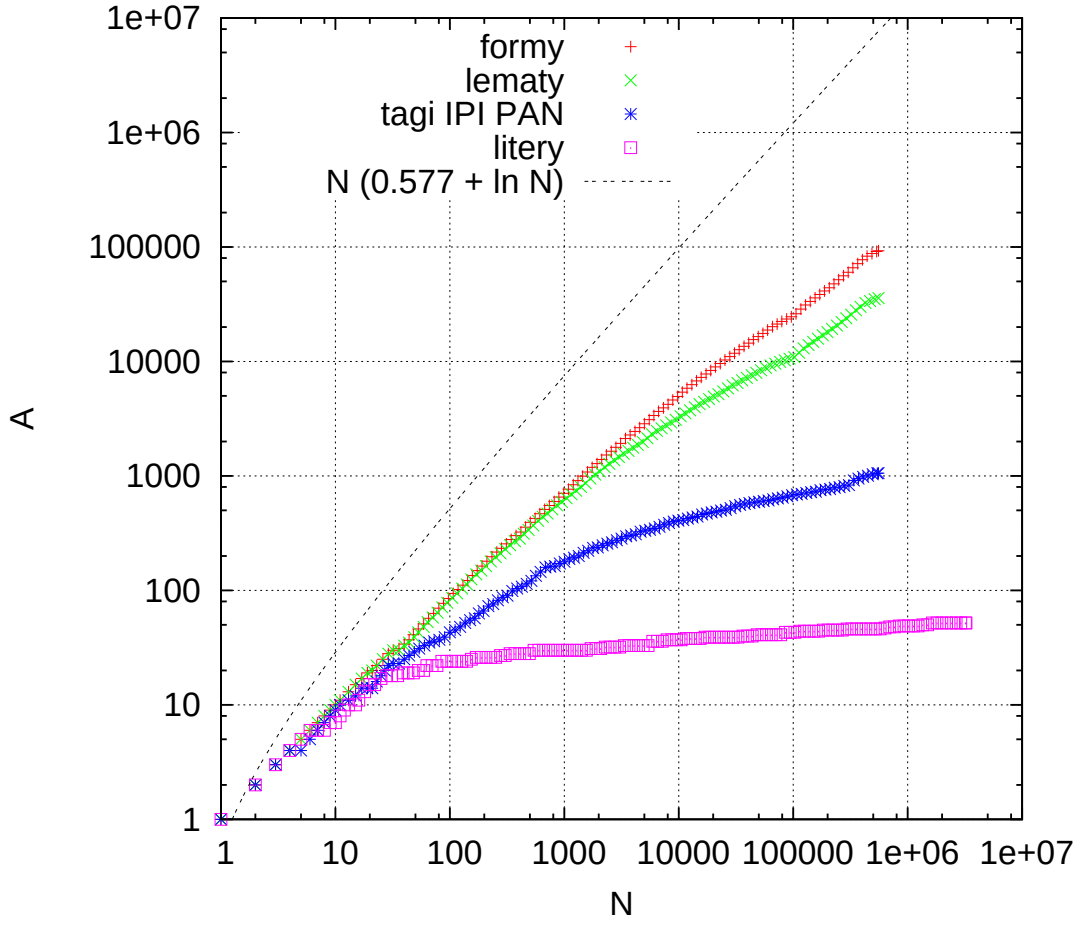
Stąd $\mathbf{N} \leq \mathbf{A}(\gamma + \ln \mathbf{N})$, czyli

$$\mathbf{A} \geq \frac{\mathbf{N}}{\gamma + \ln \mathbf{N}}.$$

Jednakże liczba różnych słów w tekście rośnie znacznie wolniej.

Wykres w skali logarytmicznej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Przybliżony wzór empiryczny

A — liczba różnych słów (typów) w tekście
N — liczba wszystkich słów (okazów) w tekście

Gustav Herdan (1964), *Quantitative Linguistics*, Butterworths & Co., zaproponował przybliżony wzór (zwaną też prawem Heapsa):

$$\mathbf{A} \approx \mathbf{a}\mathbf{N}^{\mathbf{C}}.$$

Dla korpusu SFPW mamy $\mathbf{C} \approx \mathbf{2/3}$ dla $\mathbf{N} > \mathbf{1000}$.

- H. S. Heaps (1978), *Information Retrieval—Computational and Theoretical Aspects*, Academic Press.
- Gejza Wimmer, Gabriel Altmann (1999), *On Vocabulary Richness*, Journal of Quantitative Linguistics 6:1–9.

Polski akcent

Wcześniejsi autorzy proponowali $\mathbf{C = 1/2}$, czyli $\mathbf{A \approx a\sqrt{N}}$:

- Władysław Kuraszkiewicz, Józef Łukaszewicz (1951), *Ilość różnych wyrazów w zależności od długości tekstu*, Pamiętnik Literacki 42(1):168–182.
- Pierre Guiraud (1954), *Les caractères statistiques du vocabulaire*, Presses Universitaires de France.

Rozkłady LNRE

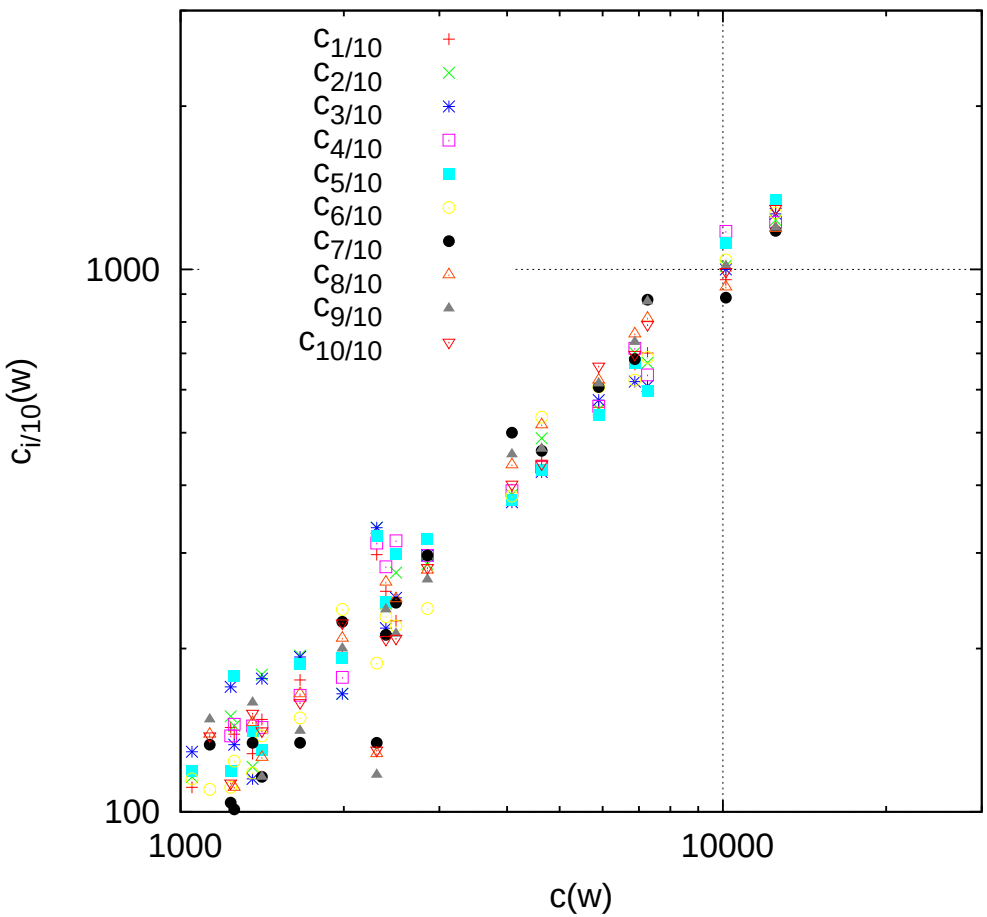
Prawo Zipfa i potęgowa zależność liczby różnych słów od długości tekstu zainspirowała sformułowanie pojęcia rozkładów LNRE (large number of rare events) w rachunku prawdopodobieństwa.

- Estate Khmaladze (1987), *The statistical analysis of large number of rare events*, Technical Report MS-R8804, Dept. of Mathematical Statistics, CWI.
- Harald Baayen (2001), *Word frequency distributions*, Kluwer Academic Publishers.

Częstości w całym tekście a częstości we fragmentach

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

$\mathbf{c_{i/10}(w)}$ — częstość słowa **w** w **i**-tym z **10** fragmentów.



Twierdzenie

Jeżeli

1. prawo Zipfa obowiązuje dla każdego tekstu długości $\mathbf{N'}$ słów,
2. w każdym tekście długości $\mathbf{N'}$ słów najczęstszym słowem jest to samo słowo,

to prawo Zipfa nie obowiązuje dla tekstów długości $\mathbf{N} \gg \mathbf{N'}$.

Dlaczego?

Niech $\mathbf{A'}$ będzie liczbą różnych słów w dowolnym tekście długości $\mathbf{N'}$. Liczba $\mathbf{A'}$ jest także częstością najczęstszego słowa w dowolnym z tych tekstów.

Tekst długości $\mathbf{N}$ podzielmy na $\mathbf{N/N'}$ fragmentów długości $\mathbf{N'}$.

W każdym fragmencie najczęstszym słowem jest to samo słowo. A zatem jest ono najczęstszym słowem w całym tekście i jego częstość wynosi $\mathbf{A'N/N'}$.

Inaczej mówiąc, w dowolnym tekście długości $\mathbf{N} \gg \mathbf{N'}$ wystąpienia najczęstszego słowa to $\mathbf{A'/N'}$ wszystkich okazów.

Sprowadzenie do niedorzeczności

Jednakże, jeżeli prawo Zipfa obowiązuje dla tekstów długości $\mathbf{N}$, to z nierówności $\mathbf{A(\gamma - 1 + \ln A) \leq N \leq A(\gamma + \ln N)}$ wynika, że

$$\frac{\mathbf{c(1)}}{\mathbf{N}} \approx \frac{\mathbf{A}}{\mathbf{N}} \leq \frac{\mathbf{1}}{\mathbf{\gamma - 1 + \ln A}} \leq \frac{\mathbf{1}}{\mathbf{\gamma - 1 + \ln \frac{N}{\gamma + \ln N}}} \xrightarrow[\mathbf{N \to \infty}]{} \mathbf{0.}$$

Wzór Mandelbrota

Benoit B. Mandelbrot (1954), *Structure formelle des textes et communication*, Word 10:1–27, zaproponował wzór

$$\mathbf{c(w)} \approx \left\lfloor \left(\frac{\boldsymbol{\rho} + \mathbf{A}}{\boldsymbol{\rho} + \mathbf{r(w)}} \right)^{\mathbf{B}} \right\rfloor, \quad \mathbf{B} > \mathbf{1}.$$

W zamyśle parametry $\boldsymbol{\rho}$ i $\mathbf{B}$ miały nie zależeć od tekstu.

Motywacje wzoru Mandelbrota

1. uniknięcie przeszacowania częstości najczęstszych słów,
2. dla stałego $\mathbf{B > 1}$ stosunek $\mathbf{c(1)/N}$ dąży do stałej,

$$\frac{\mathbf{c(1)}}{\mathbf{N}} \xrightarrow[\mathbf{N\to\infty}]{} \pi\mathbf{(1) > 0},$$

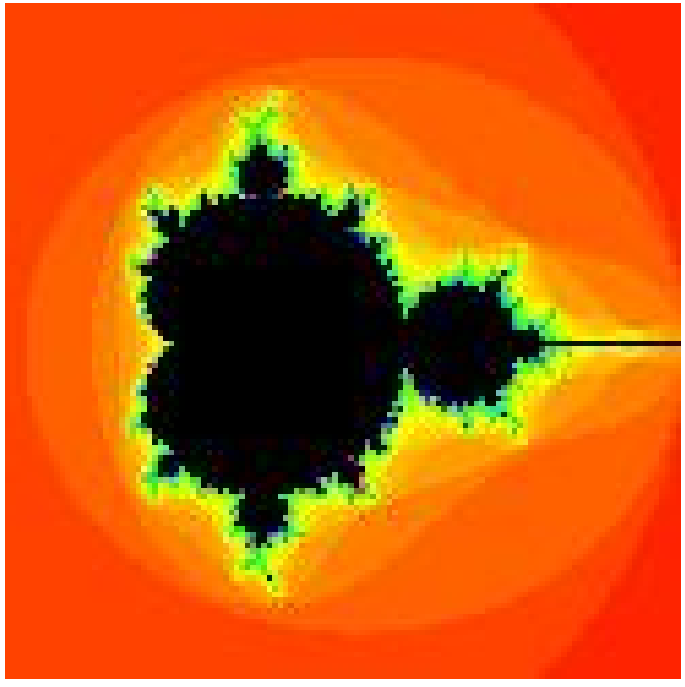
3. potęgowa zależność liczby typów $\mathbf{A}$ od liczby okazów $\mathbf{N}$,

$$\mathbf{A \approx aN^C - \rho}, \quad \mathbf{C = 1/B}, \quad \mathbf{a = (\rho + 1)\pi(1)^C},$$

- A. Kornai (2002), *How many words are there?*, Glottometrics 4:61–86.
- D.C. van Leijenhorst, Th.P. van der Weide (2005), *A formal derivation of Heaps' Law*, Information Sciences 170:263-272.

4. wzór obowiązuje dla prostego probabilistycznego modelu tekstu (tzw. modelu "małpy przy klawiaturze").

Benoit B. Mandelbrot (1924-)

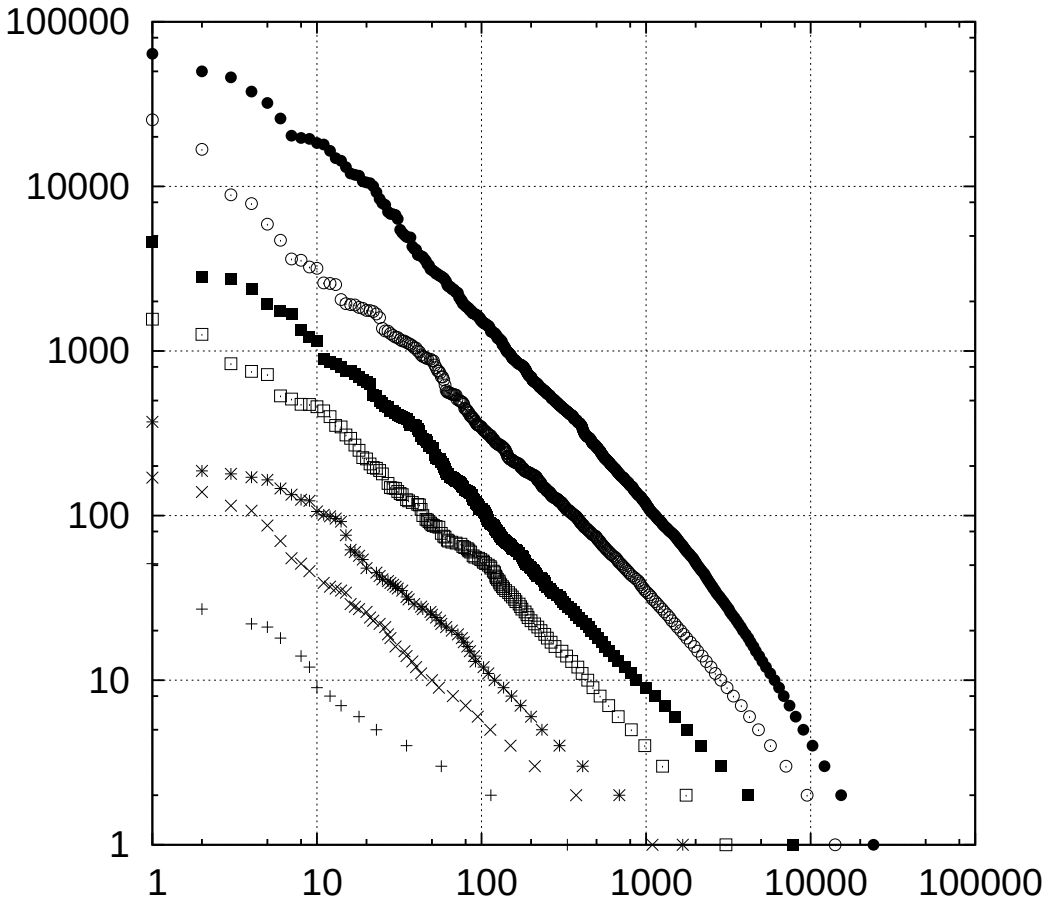


A rzeczywistość skrzeczy

Wzór zaproponowany przez Mandelbrota jest zbyt prosty:

1. Optymalne wartości parametrów $\mathbf{B}$ i ρ zależą od tekstu.
2. Dla tekstów krótkich $\mathbf{B} < \mathbf{1}$, dla długich zaś $\mathbf{B} > \mathbf{1}$.
3. Niesposób tak dobrać parametrów, aby uzyskać dobre dopasowanie do całego wykresu ranga-częstość oraz do wykresu liczba typów-liczba okazów.

Wykresy ranga-częstość dla tekstów w j. angielskim



The Complete Memoirs,
N = 1 262 287;
The Descent of Man,
N = 308 171;
Erewhon,
N = 84 717;
Through the Looking-Glass,
N = 31 055;
Adventure of the Red Circle,
N = 7407;
A Modest Proposal,
N = 3427;
Peach Blossom Shangri-la,
N = 735.

Literatura nt. zależności parametrów od długości tekstu

- Ju. K. Orlov (1982), *Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses?*, [w:] Ju. K. Orlov, M. G. Boroda, I. Š. Nadarejšvili, Sprache, Text, Kunst. Quantitative Analysen, 1–55, Dr. N. Brockenmeyer.
- Harald Baayen (2001), *Word frequency distributions*, Kluwer Academic Publishers.
- Łukasz Dębowski (2002), *Zipf's law against the text size: A half-rational model*, Glottometrics, 4:49–60.

Problem ekstrapolacji

Dotychczasowe ilustracje prawa Zipfa:

- Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej* — **555 073** okazów,
- Siedem jednolitych tekstów literackich w języku angielskim — od **735** do **1 262 287** okazów.

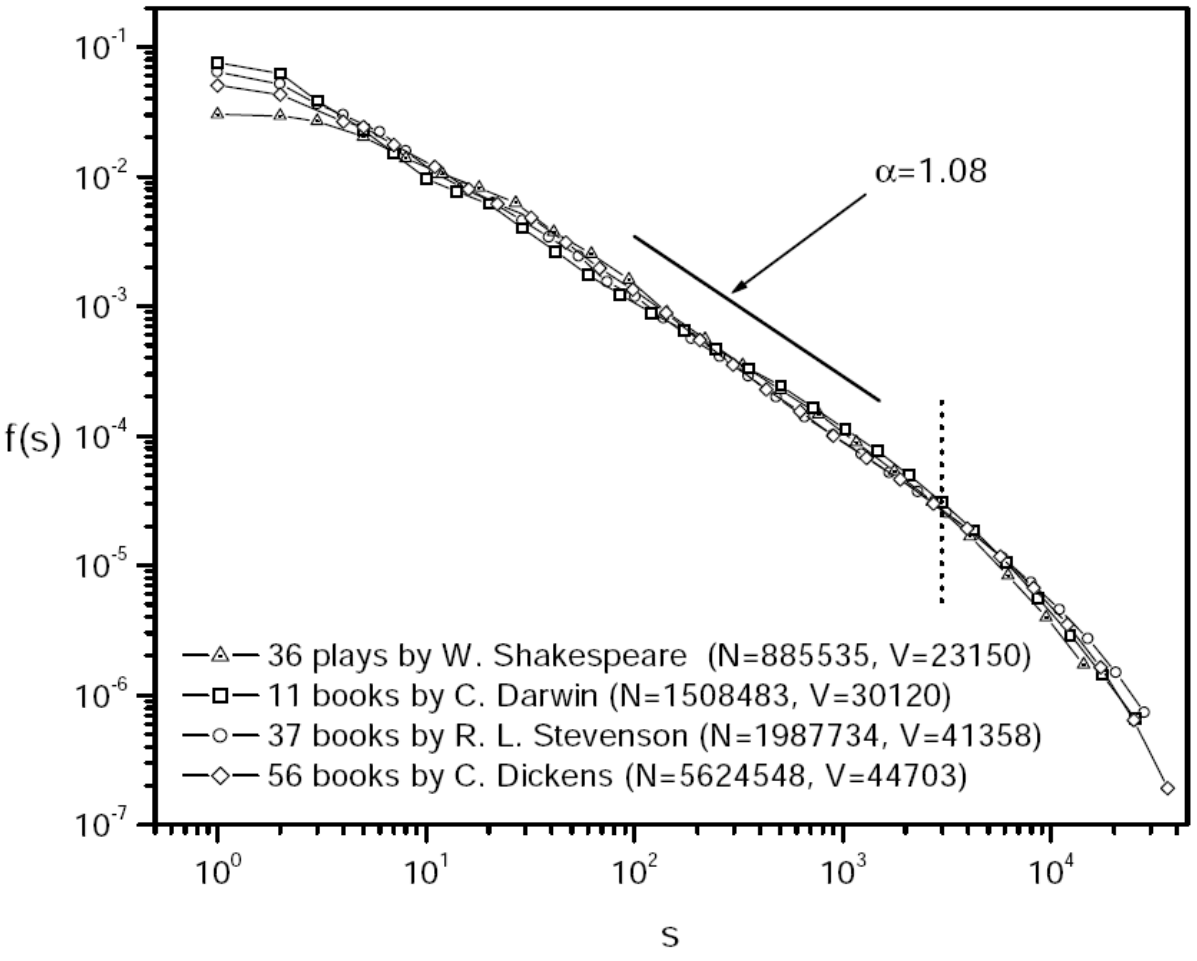
Czy dla korpusów liczących 100–1000 mln okazów przybliżenie Mandelbrota jest nadal sensowne?

Ramon Ferrer i Cancho oraz Marcelo A. Montemurro

- R. Ferrer i Cancho, R. V. Solé (2001), *Two regimes in the frequency of words and the origin of complex lexicons*, Journal of Quantitative Linguistics 8:165–173.
- M. A. Montemurro, D. H. Zanette (2002), *New perspectives on Zipf's law in linguistics: From single texts to large corpora*, Glottometrics 4:86–98.

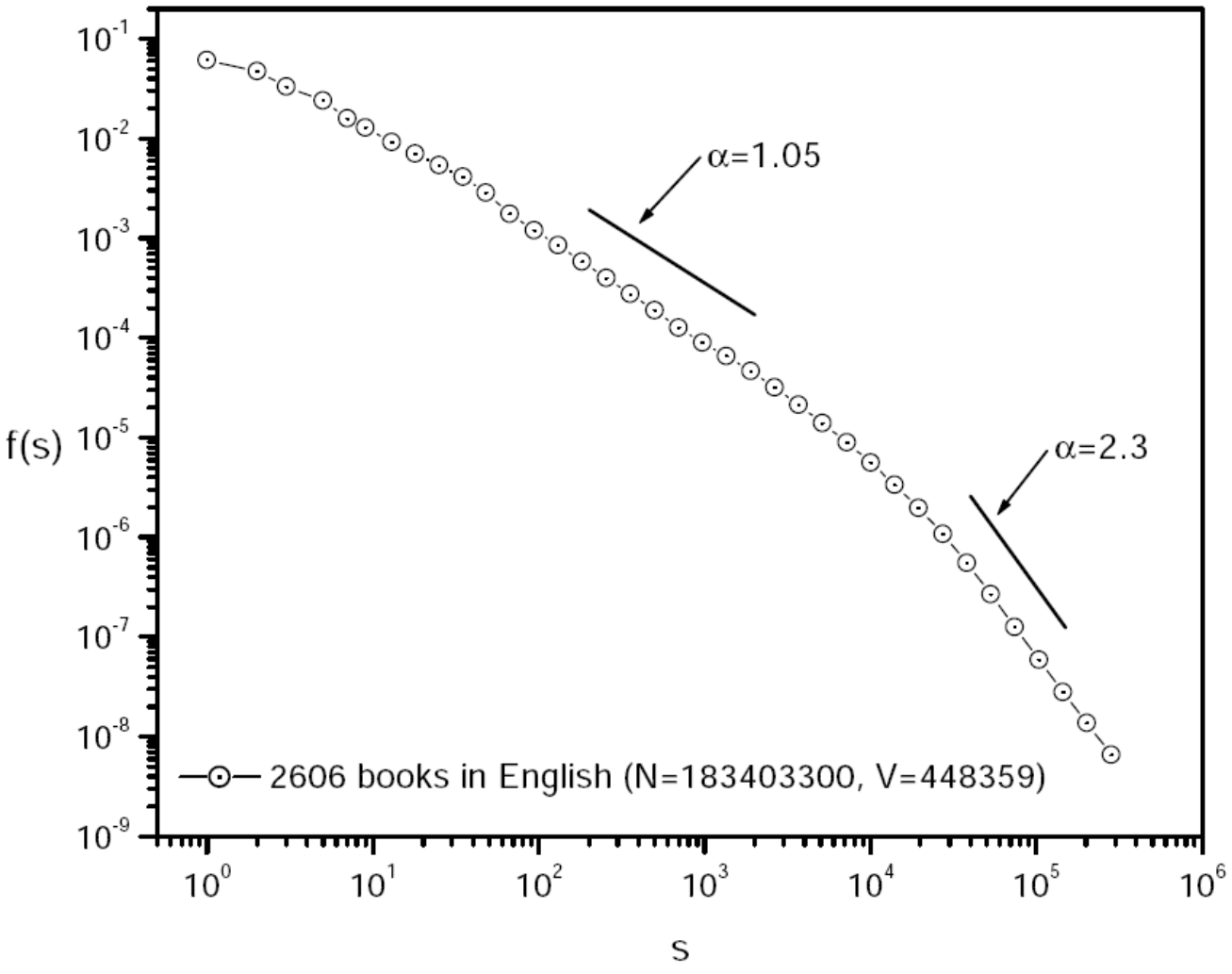
Wykres ranga-częstość względna

Cztery korpusy dzieł pojedynczych autorów (**N** $\approx$ **1** mln)



Wykres ranga-częstość względna

Jeden korpus dzieł wielu autorów (**N ≈ 100** mln)



Trzy klasy słów na wykresie ranga-częstość

(I) słowa najczestsze:
— poczatkowe spłaszczenie wykresu

(II) słowa średniej częstości:
— częstość odwrotnie proporcjonalna do rangi (prawo Zipfa)

(IIIa) słowa rzadkie dla korpusu dzieł jednego autora:
— wykładniczy zanik czestości w funkcji rangi

(IIIb) słowa rzadkie dla korpusu dzieł wielu autorów:
— częstość odwrotnie proporcjonalna do kwadratu (?) rangi

Interpretacja?

Słownictwo aktywnie używane przez pojedynczego człowieka jest mocno ograniczone, ale silnie zindywidualizowane.

Szacowanie rozmiarów aktywnego leksykonu z poprzednich wykresów jest ryzykowne.

Podział słownictwa na wspomniane trzy klasy może zależeć od korpusu, w tym jego rozmiaru.

Prawo Zipfa dla **n**-tek (teksty długości **N** $\approx$ **100** mln)

Sporządźmy listy frekwencyjne dla wszystkich **n**-tek słów (względnie liter lub sylab) pojawiających się w korpusie, gdzie $\mathbf{n = 1, 2, 3, \ldots}$, i połączmy te listy w jedną listę frekwencyjną.

Dla połączonej listy frekwencyjnej spełnione jest prawo Zipfa postaci $\mathbf{c(r) \propto 1/r}$ dla typów o najmniejszej częstości.

- Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, F. J. Smith (2003), *Extension of Zipf's Law to Word and Character* **N***-grams for English and Chinese*, Computational Linguistics and Chinese Language Processing, 8(1):77–102.

Czy dokładne prawo Zipfa dla **n**-tek liter jest lepszą charakterystyką probabilistycznych własności języka niż przybliżone prawo Zipfa dla słów?

Podsumowanie

- Odstępstwa wykresu ranga-częstość od prawa Zipfa są systematyczne. Świadczą o tym:
  - wzrost liczby różnych słów w funkcji długości tekstu,
  - proporcjonalność częstości najczęstszych słów do dł. tekstu.
- Pewną próbą opisu odstępstw jest wzór Mandelbrota, ale rzeczywistość jest znacznie bardziej skomplikowana.
- Wykres ranga-częstość zależy od rodzaju tekstów, dla których został sporządzony.