Prawo Zipfa
zjawiska (I)

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Ilościowa analiza tekstu?

Teksty tworzone przez ludzi mają złożoną strukturę formalną:

- na poziomie morfologii słów,
- na poziomie składni zdań,
- na poziomie narracji tekstu.

Czy teksty te wykazują także pewne prawidłowości ilościowe?

Zamiast spekulować, przeprowadźmy prosty eksperyment.

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
...	...	...

Przemilczana kwestia definicji słowa

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
...	...	...
14	2292	W
15	2156	em
16	1968	co
17	1888	Nie
18	1852	od
19	1789	tym
20	1758	tak
21	1711	po
22	1631	już
23	1575	by
...	...	...

Pierwsze powtórzenia częstości

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
...	...	...
206	214	sto
207	214	kto
208	214	kilka
209	214	człowiek
210	214	ciągu
211	213	jeśli
212	212	czas
213	210	ludzie
214	209	niej
215	208	takich
...	...	...

Hapax legomena

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
...	...	...
38419	2	abażury
38420	2	Aaa
38421	1	żyznej
...	...	...
55267	1	skandowali
...	...	...
66665	1	pantofagiem
...	...	...
79660	1	kijowskiego
...	...	...
92963	1	aa

Truizm

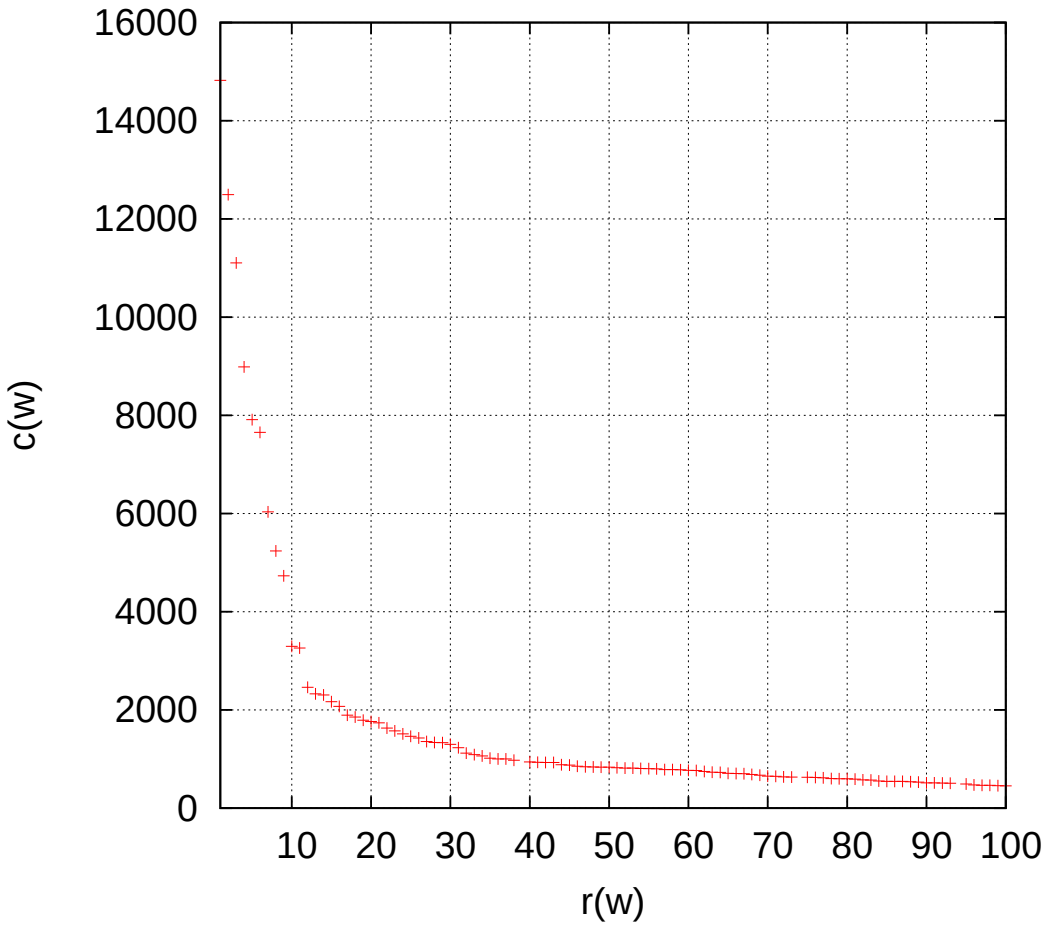
Definicja rangi słowa:

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

Im większa ranga, tym mniejsza częstość.

Wykres w skali linowej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

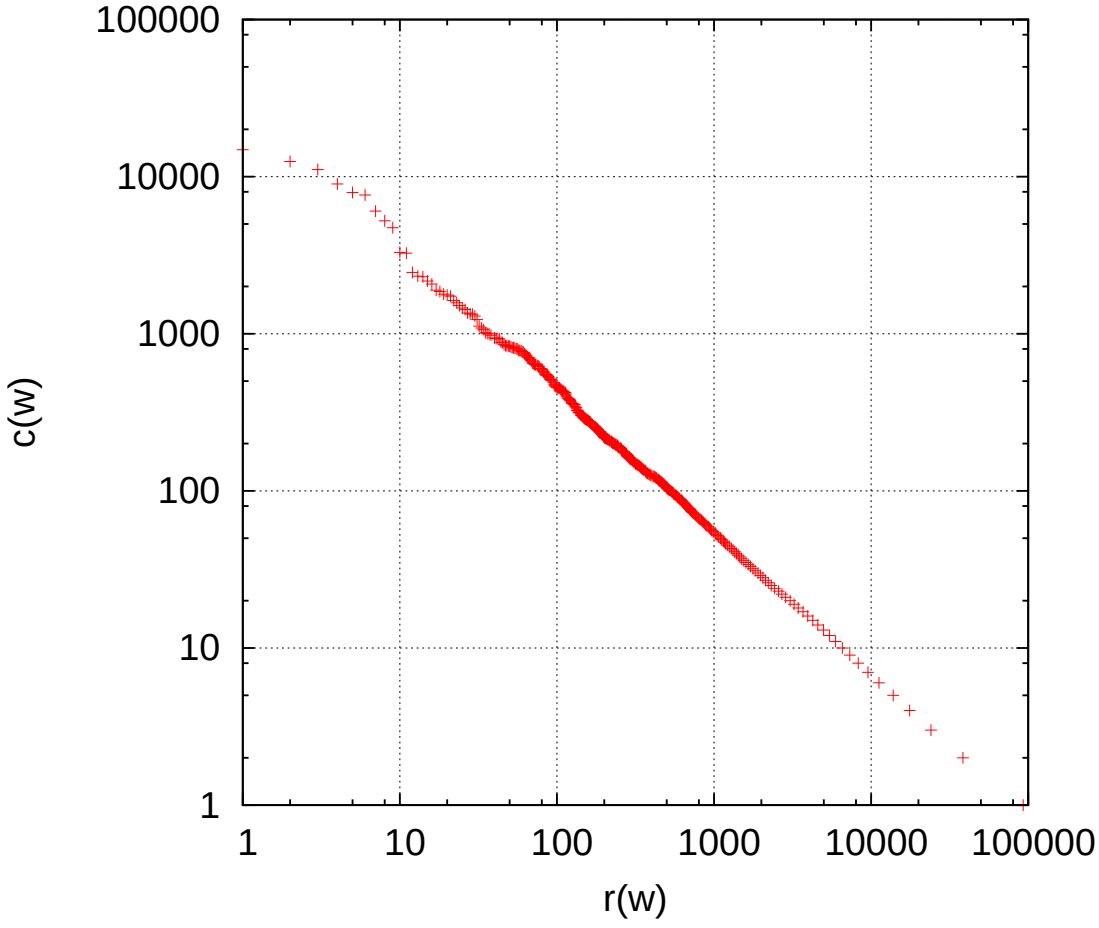


Zastrzeżenie

Definicja nie mówi, jak szybko ranga maleje z częstością.

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:

Dla każdego słowa w tekście iloczyn rangi i częstości jest stały,

$$\mathbf{r(w) \cdot c(w) \approx A.}$$

Równoważnie:

Częstość słowa jest odwrotnie proporcjonalna do jego rangi,

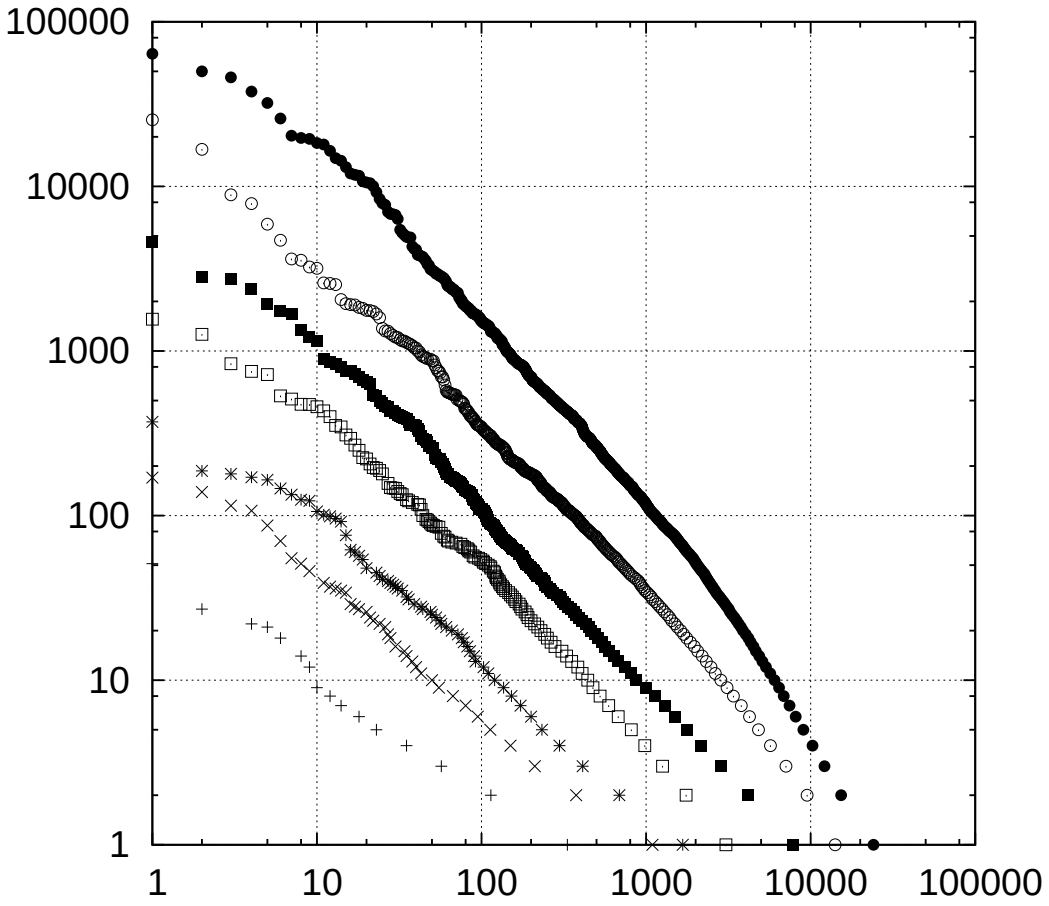
$$\mathbf{c(w) \approx \frac{A}{r(w)}.}$$

Ale czy to aby takie niezwykłe?

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

Wykresy dla listy rangowej liter wyglądają inaczej!

Rangi i częstości słów w tekstach 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.

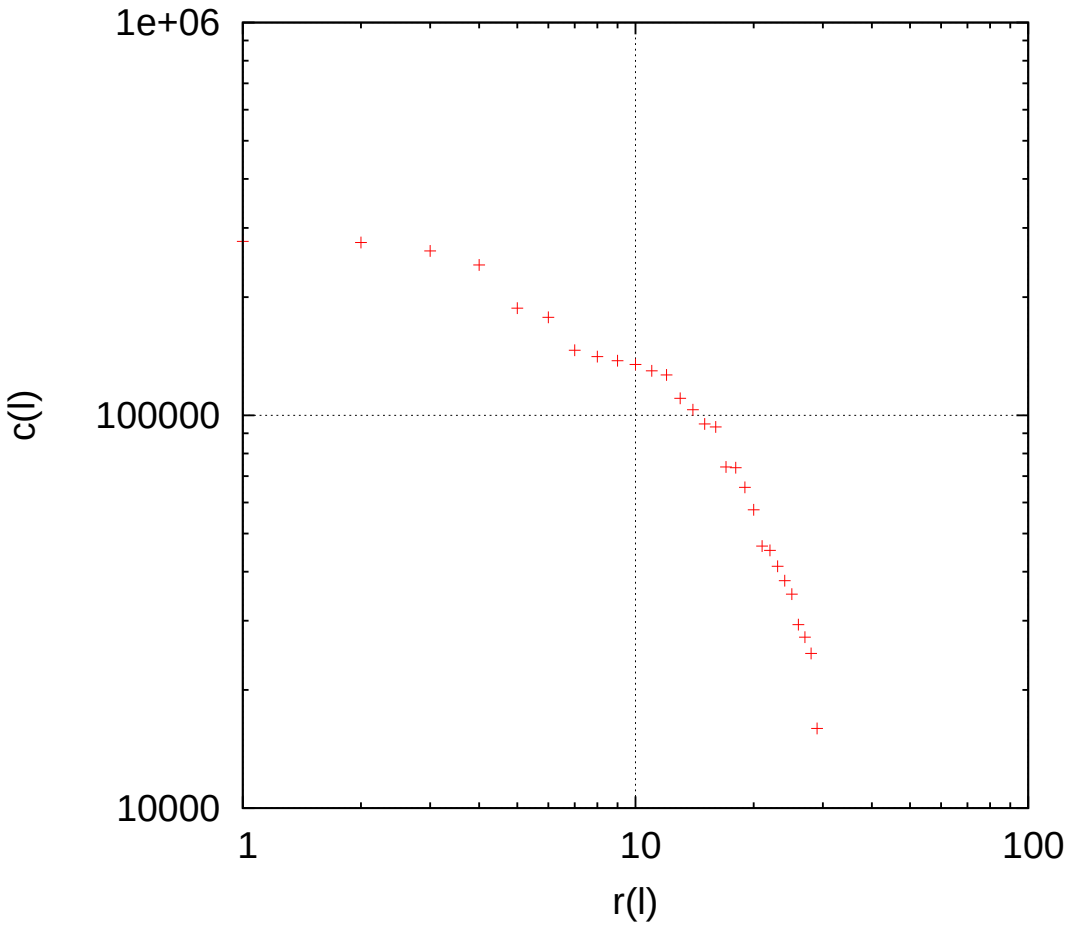
Lista rangowa liter

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(l)**	częstość **c(l)**	litera **l**
1	276580	i
2	274921	a
3	261696	e
4	241076	o
5	187131	z
6	177350	n
7	146198	w
8	140830	r
9	137425	s
...	...	...

Wykres ranga-częstość dla liter

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Prawo Zipfa a lingwistyka kwantytatywna

Prawo Zipfa jest najsłynniejszym ilościowym prawem językowym.

Badaniem ilościowych praw językowych zajmuje się lingwistyka kwantytatywna.

Rozkłady ranga-częstość podobne do prawa Zipfa pojawiają się także poza lingwistyką:

- rozkład cytowań artykułów naukowych (prawo Lotki),
- rozkład dochodów ludności (prawo Pareto, zasada 80/20),
- rozkład wielkości miast (prawo Gibrata).

Klasycy prawa Zipfa

- Vilfredo Pareto (1896), *Cours d'économie politique*.
- Jean-Baptiste Estoup (1916), *Les Gammes Stenographiques*.
- Alfred J. Lotka (1926), *The frequency distribution of scientific productivity*, Journal of the Washington Academy of Sciences, 16(12):317–324.
- George K. Zipf (1935), *Psycho-Biology of Language. An Introduction to Dynamic Philology*.
- George K. Zipf (1949), *Human Behavior and the Principle of Least Effort*.

George Kingsley Zipf (1902-1950)

Zipf was the kind of man who would take roses apart to count their petals.

— George A. Miller

Współczesność

Odkryto kilka rozumowań probabilistycznych, które satysfakcjonująco objaśniają pojawianie się prawa Zipfa w niektórych kontekstach pozalingwistycznych.

Nadal brak jest dopracowanego modelu matematycznego, który wyjaśniałby pojawianie się prawa Zipfa w odniesieniu do rozkładu słów w tekstach tworzonych przez ludzi.

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/

Formalistyczna uwaga

Wykres ranga-częstość zależy od tego,
czego częstości i rangi rozpatrujemy.

Okazy i typy

Okazy to słowa (obiekty) pojawiające się w tekście (danych).

Typy to słowa (obiekty) pojawiające się na liście rangowej.

Procedura liczenia częstości typów:

1. Najpierw dzielimy tekst na okazy.
2. Częstość danego typu to liczba okazów tego typu.

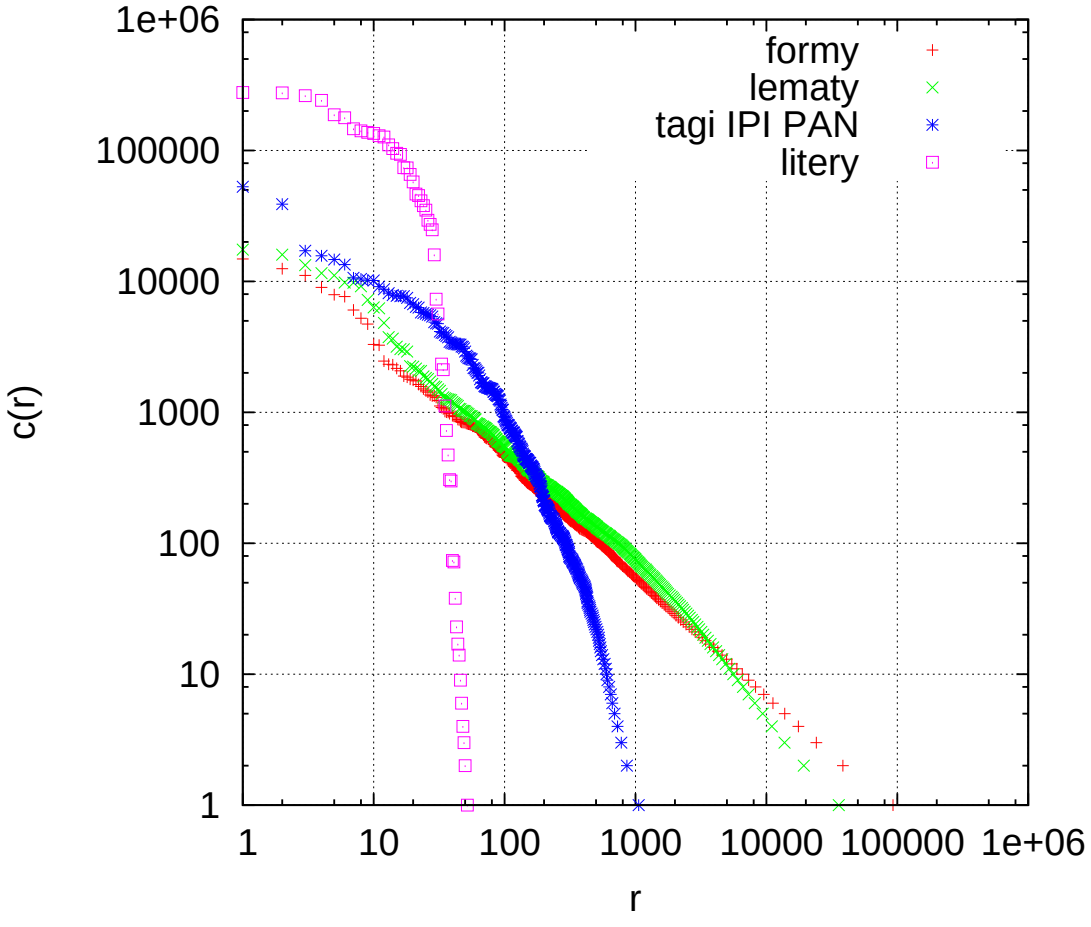
Liczba okazów to długość tekstu w słowach.
Liczba typów to liczba różnych słów w tekście.

Okazy i typy

- Gdzie w tekście znajdują się okazy? Możliwości:
  1. `Nauczyliśmy się szukać wszędzie przyczyn...`
     `<---------> <-> <----> <------> <------>   ;`
  2. `Nauczyliśmy się szukać wszędzie przyczyn...`
     `<------><-> <-> <----> <------> <------>   .`
- Jakie okazy należą do danego typu? Możliwości:
  1. `jest -> jest, Jest -> Jest, są -> są, Są -> Są ;`
  2. `jest -> jest, Jest -> jest, są -> są, Są -> są ;`
  3. `jest -> być, Jest -> być, są -> być, Są -> być .`

Wykresy ranga-częstość dla różnych obiektów

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Pytanie otwarte

Czy istnieją takie definicje okazu i typu,
dla których prawo Zipfa można wydedukować?

Częstość częstości

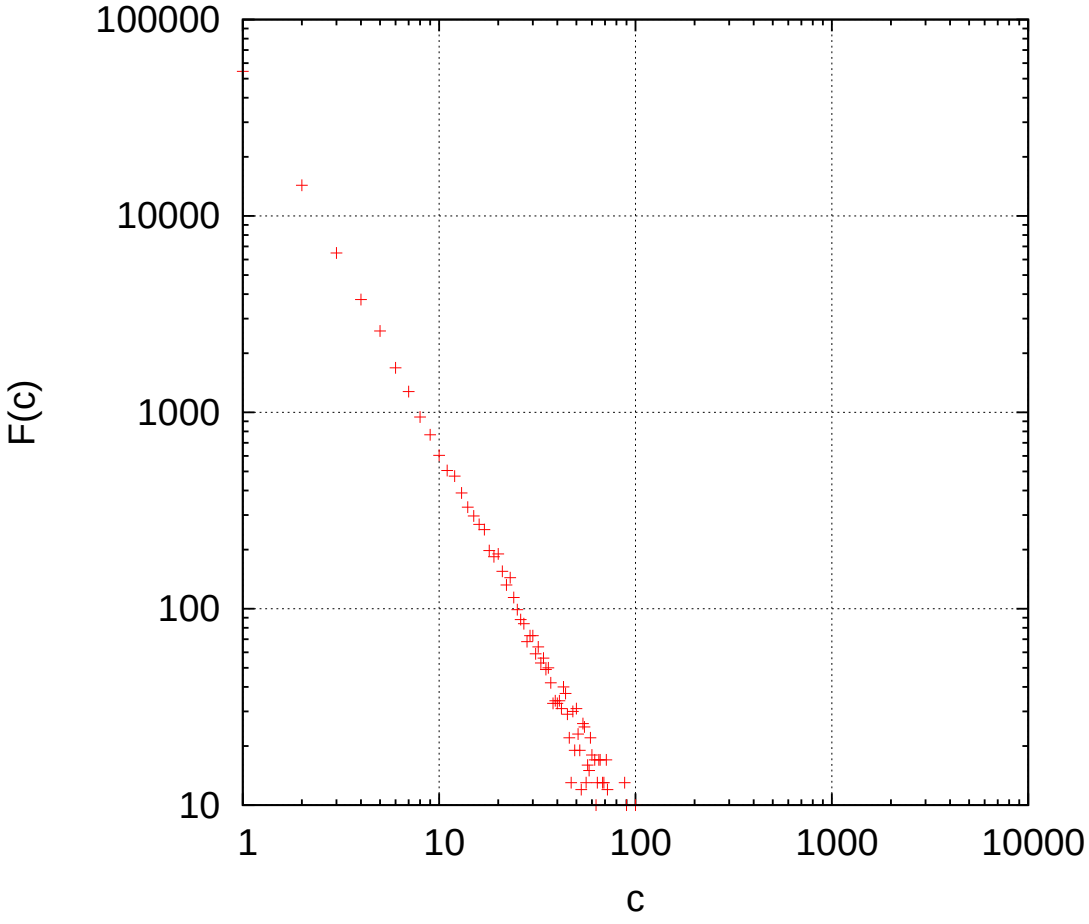
Częstość częstości to liczba różnych słów (typów) o danej częstości.

F(c) — liczba typów o częstości **c**.

- **F(1)** — liczba słów występujących w tekście 1 raz (hapaksy),
- **F(2)** — liczba słów występujących w tekście 2 razy.
- **F(2)** — liczba słów występujących w tekście 3 razy, ... itd.

Wykres częstości częstości w skali logarytmicznej

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Prawo Lotki

Jeżeli $\mathbf{F(c_1) = 10F(c_2)}$, to $\mathbf{100c_1 = c_2}$.

Równoważnie:

Iloczyn częstość częstości $\mathbf{F(c)}$ i kwadratu częstości $\mathbf{c}$ jest stały.

Dokładniej,

$$\mathbf{F(c) \approx \frac{A}{c(c+1)}},$$

gdzie $\mathbf{A}$ to liczba różnych słów pojawiających się w tekście.

Wniosek:

Połowa słownictwa każdego tekstu to hapaksy!

Ile jest hapaksów naprawdę?

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
...	...	...
24076	3	[&][#]
24077	2	żywotnych
...	...	...
38420	2	Aaa
38421	1	życznej
...	...	...
92963	1	aa

$$\mathbf{(92963 - 38420)/92963 = 0,587... \approx 1/2}$$

$$\mathbf{(38420 - 24076)/92963 = 0,154... \approx 1/6}$$

Uwaga

Prawo Lotki wynika z prawa Zipfa!

Wyprowadzenie prawa Lotki

F(c) — liczba typów o częstości **c**
c(r) — częstość typu o randze **r**
r(c) — największa ranga typu o częstości **c**

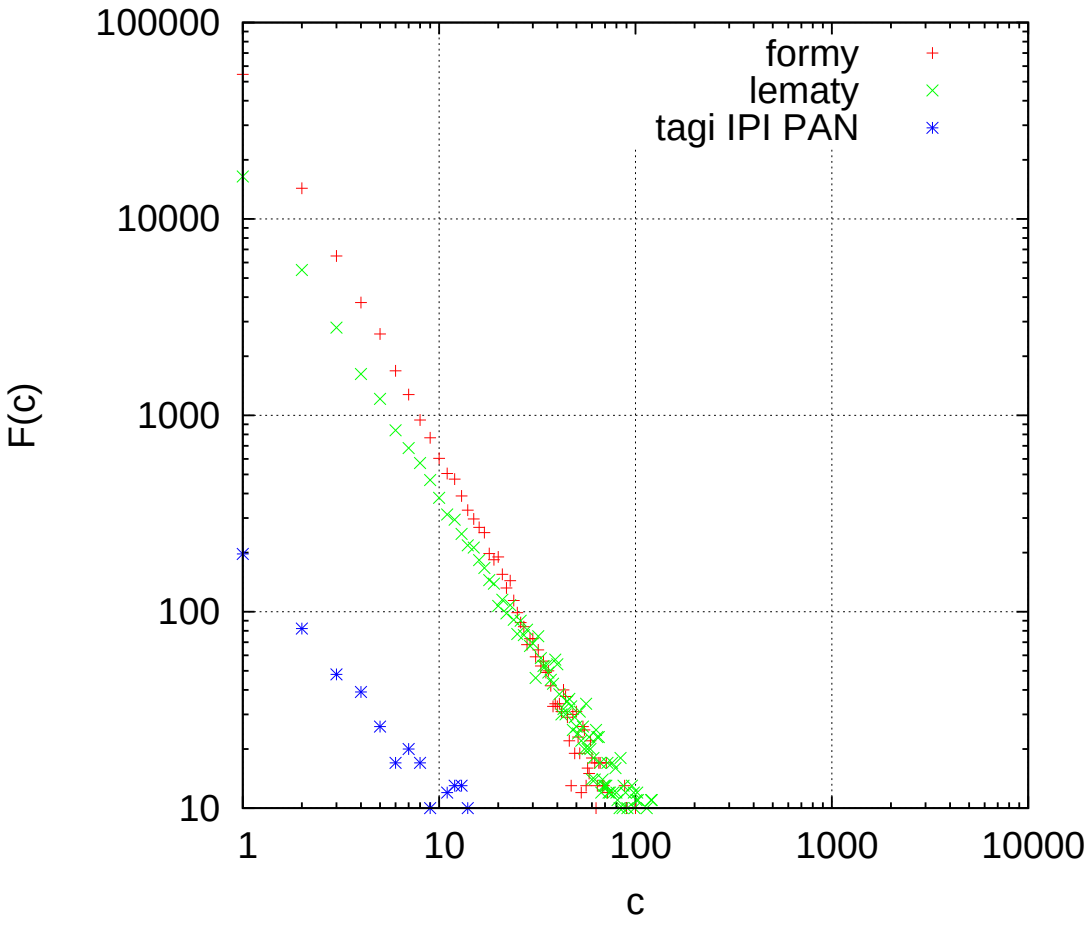
$$\mathbf{c(r)} = \left\lfloor \frac{\mathbf{A}}{\mathbf{r}} \right\rfloor \qquad \text{(prawo Zipfa)}$$

Z prawa Zipfa, $\mathbf{c(r) = 0}$ wtedy i tylko wtedy, gdy $\mathbf{r > A}$.
A zatem **A** to liczba różnych słów pojawiających się w tekście.
Mamy $\mathbf{r(c) = A/c}$ oraz

$$\mathbf{F(c) = r(c) - r(c+1) = \frac{A}{c} - \frac{A}{c+1}}$$
$$\mathbf{= \frac{A(c+1) - Ac}{c(c+1)} = \frac{A}{c(c+1)}.}$$

Wykresy częstości częstości dla innych obiektów

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*



Bibliometria

Prawo Lotki odkryto 10 lat wcześniej niż prawo Zipfa.

W dodatku nie w odniesieniu do rozkładu słów w tekstach, lecz w odniesieniu do rozkładu cytowań w pracach naukowych:

Liczba autorów cytowanych **n** razy
jest odwrotnie proporcjonalna do $\mathbf{n^2}$.

Alfred J. Lotka (1926), *The frequency distribution of scientific productivity*, J. Washington Academy Sci., 16(12):317–324.

Alfred James Lotka (1880-1949)

Interesował się modelami matematycznymi w chemii, demografii i ekologii.

Urodził się we Lwowie, wyjechał do USA w 1902.

Podsumowanie

- Ranga słowa w tekście jest w przybliżeniu odwrotnie proporcjonalna do jego częstości (prawo Zipfa).
- **W konsekwencji:**
  Połowa słownictwa tekstu to hapax legomena — słowa o jednokrotnych wystąpieniach (prawo Lotki).
- Powyższe zależności obowiązują dla szerokiego zakresu tekstów w różnych językach (w tym dla niewielkich korpusów) przy dość nieostrej definicji słowa.

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