

Rola interpunkcji w lingwistycznym prawie Zipfa

Jarosław Kwapien

Instytut Fizyki Jądrowej PAN, Kraków

*Seminarium Interdyscyplinarnego Centrum Nauk Fizycznych,
Chemicznych i Medycznych*

Instytut Fizyki Jądrowej Polskiej Akademii Nauk

16 II 2017

Cechy języka naturalnego:

- Język naturalny jest systemem przekazu informacji za pomocą znaków.
- Relacjami między znakami języka kierują strukturalne reguły, zwane gramatyką.
- Język naturalny powstał na drodze ewolucji gatunku ludzkiego i jest dynamiczną strukturą podlegającą samoorganizacji.
- Język przyczynił się do szybkiego rozwoju kory mózgowej u ludzi.
- Język mógł doprowadzić do powstania świadomości.
- Spełnia wszystkie kryteria i może być uważany za układ złożony.
- Dane empiryczne są praktycznie pozbawione błędów systematycznych.

Prosty ilościowy opis języka:

Zliczanie słów w próbce tekstu:

R – ranga słowa

F – liczba wystąpień słowa w tekście

Rozkład rangowy:

Jaki jest rozkład częstości F występowania słów w zależności od ich rangi R ?

$R = 1$ → najczęściej występujące zdarzenie

$R = 2$ → drugie najczęściej występujące zdarzenie

$R = 3$ → trzecie najczęściej występujące zdarzenie

...

$R = N$ → najrzadziej występujące zdarzenie (N – liczba zdarzeń)

1916 – obserwacja: częstotliwość słowa w jęz. francuskim jest odwrotnie proporcjonalna do jego rzędu

Jean-Baptiste Estoup, *Gammes Sténographique*

1932 – obserwacja podobnego zachowania słów w jęz. angielskim

Gordon K. Zipf, *Selective studies and the principle of relative frequency in language*, MIT Press

STATELY, PLUMP BUCK MULLIGAN CAME FROM THE STAIRHEAD, bearing a bowl of lather on which a mirror and a razor lay crossed. A yellow dressing gown, ungirdled, was sustained gently-behind him by the mild morning air. He held the bowl aloft and intoned:

- *Introibo ad altare Dei.*

Halted, he peered down the dark winding stairs and called up coarsely:

- Come up, Kinch. Come up, you fearful jesuit.

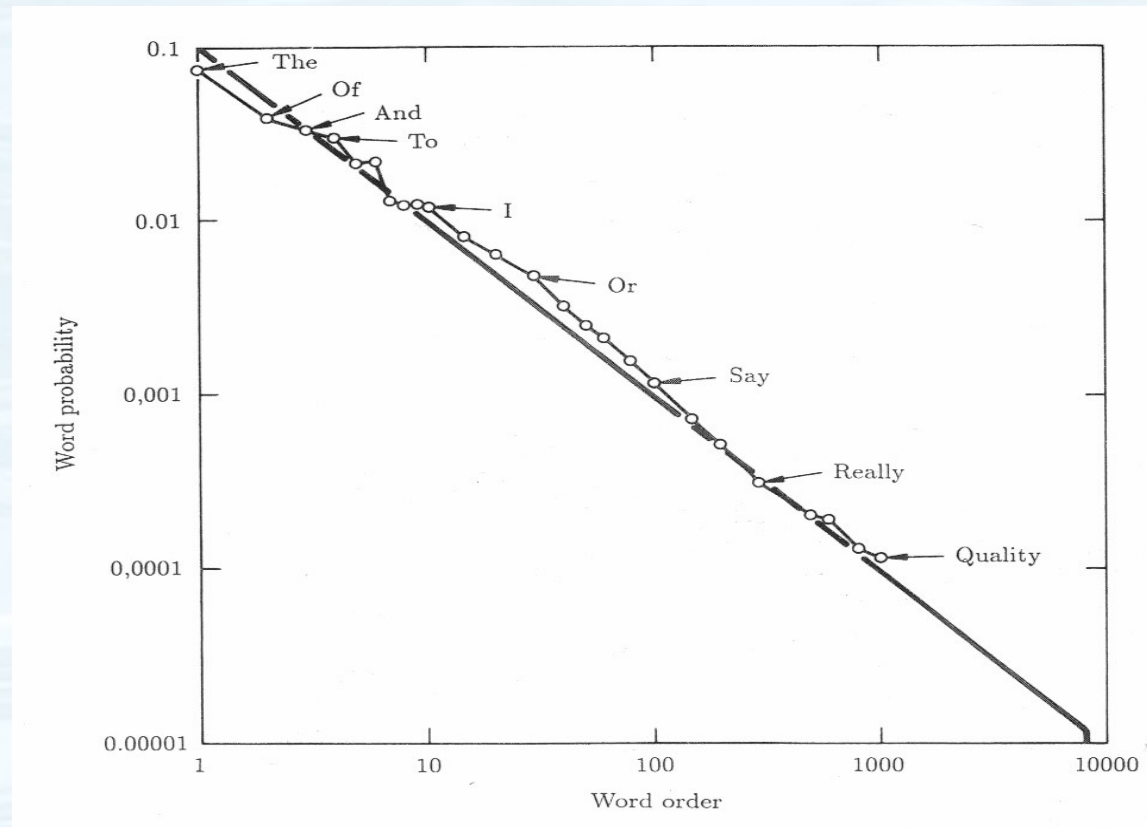
Solemnly he came forward and mounted the round gunrest. He faced about and blessed gravely thrice the tower, the surrounding country and the awaking mountains. Then, catching sight of Stephen Dedalus, he bent towards him and made rapid crosses in the air, gurgling in his throat and shaking his head. Stephen Dedalus, displeased and sleepy, leaned his arms on the top of the staircase and looked coldly at the shaking gurgling face that blessed him, equine in its length, and at the light untensured hair, grained and hued like pale oak.

James Joyce „Ulysses” Episode 1

Długość: 260 722 słów

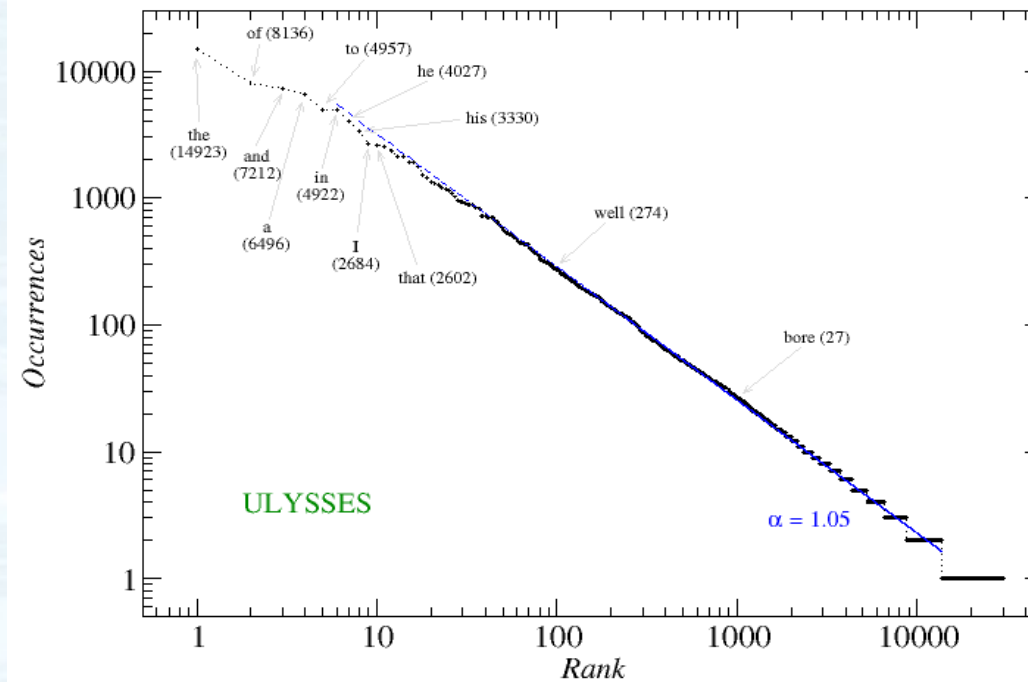
Słownik: 30 147 słów

James Joyce, "Ulysses"



Gordon K. Zipf, *Human Behavior and the Principle of Least Effort* (1949)

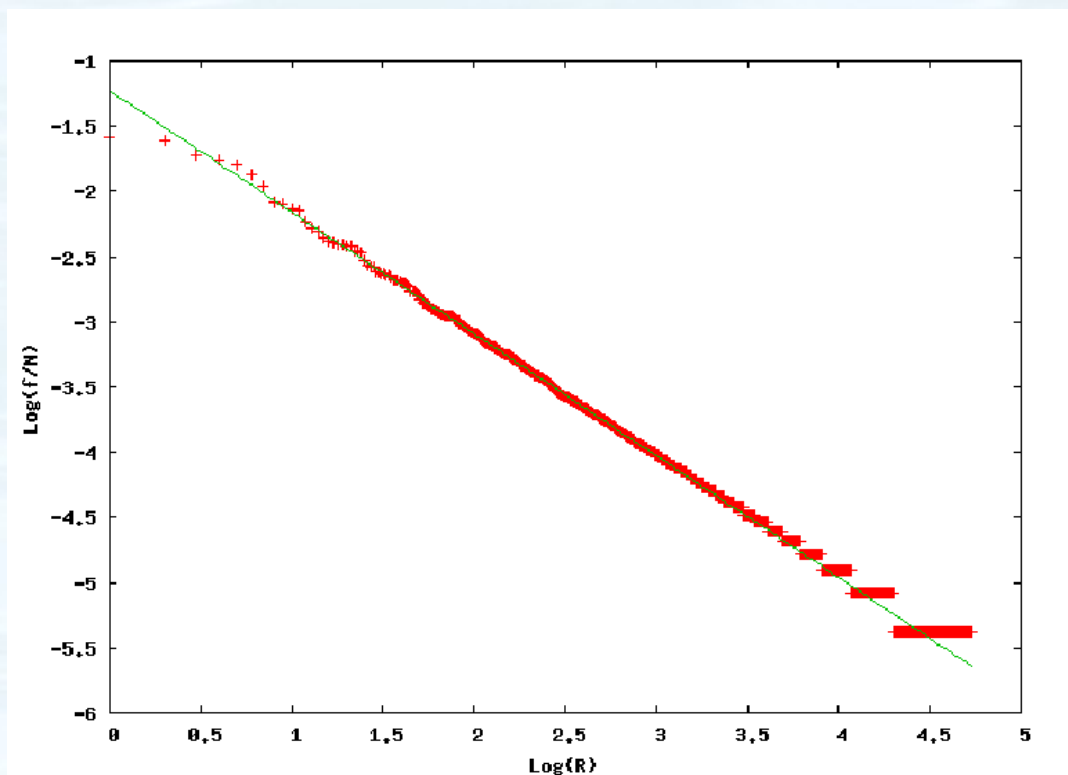
James Joyce, "Ulysses"



J.Kwapień, S. Drożdż, Physics Reports **515**, 115-226 (2012)

PRAWO ZIPFA W JĘZ. POLSKIM

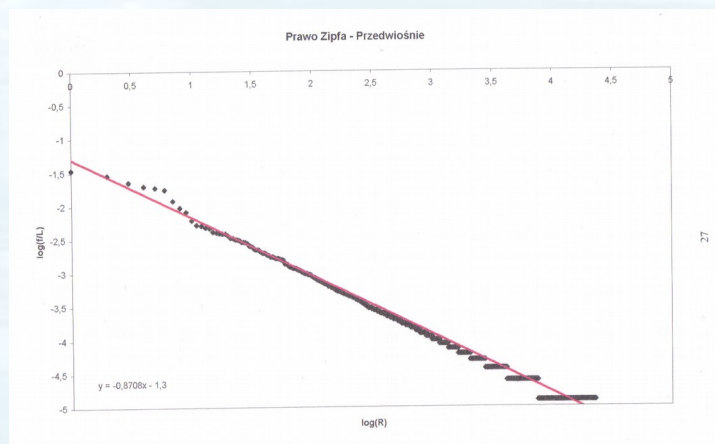
James Joyce, "Ulisses" (tłum. Maciej Słomczyński)



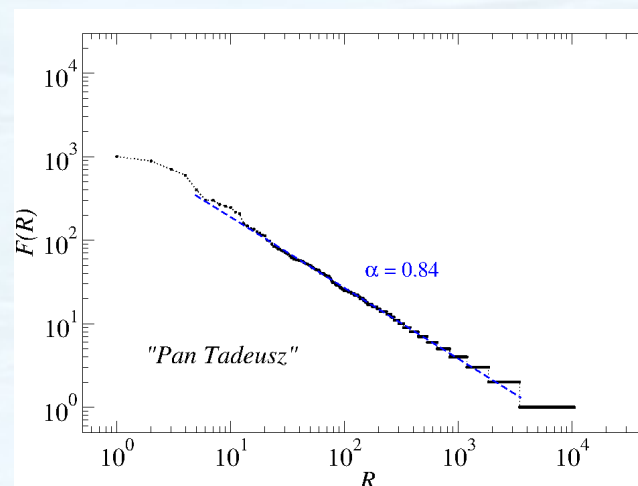
$$\alpha = 0.93$$

A. Orczyk, praca mgr (2008)

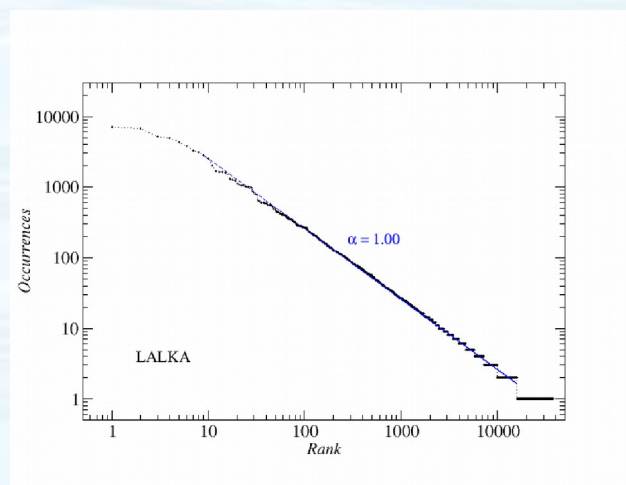
S. Żeromski, *Przedwiośnie*



A. Mickiewicz, *Pan Tadeusz*



B. Prus, *Lalka*

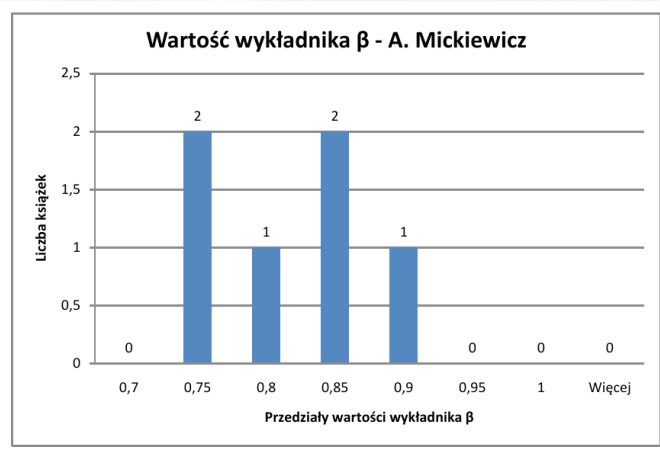


Prawo Zipfa:

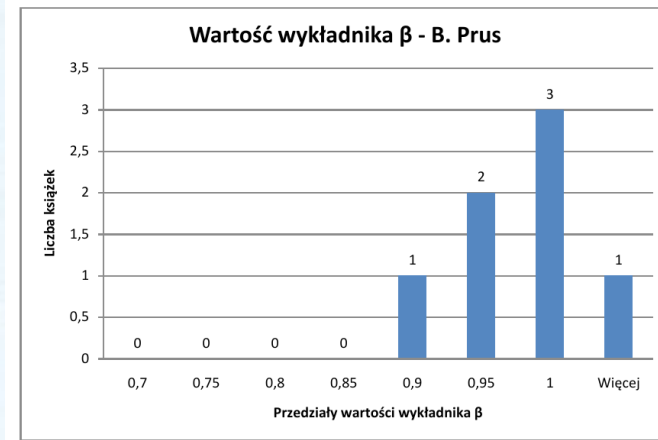
$$F(R) \sim R^{-\alpha}, \quad \alpha > 0$$

typowo: $\alpha \approx 1$

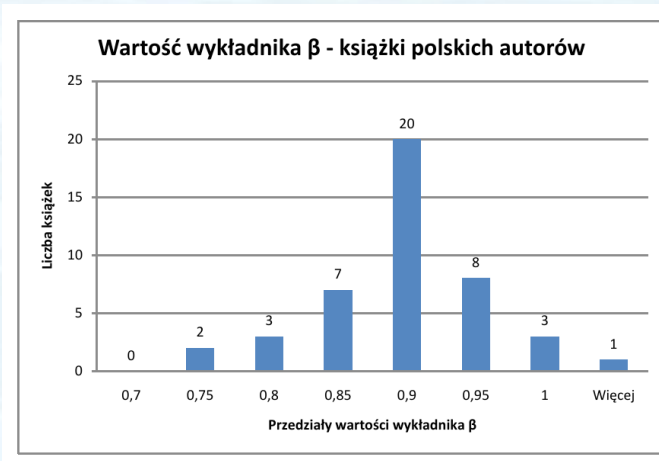
Adam Mickiewicz: $\alpha = 0.79$



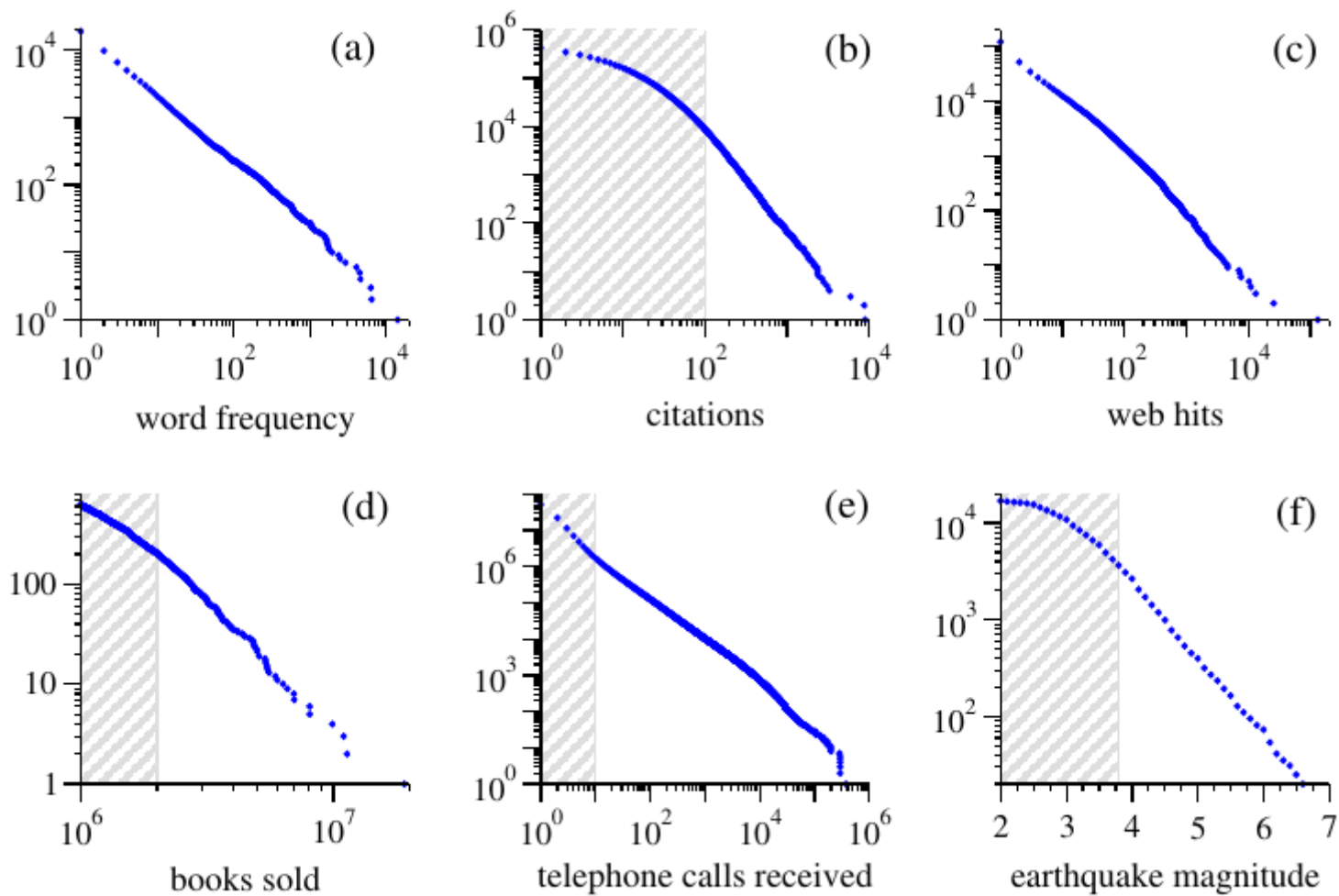
Bolesław Prus: $\alpha = 0.95$



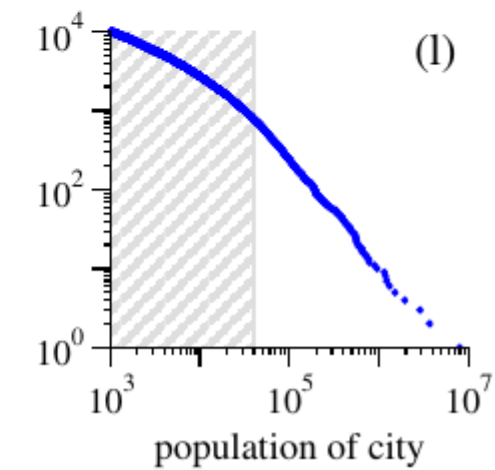
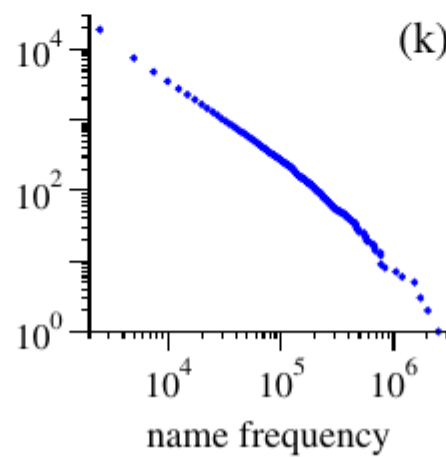
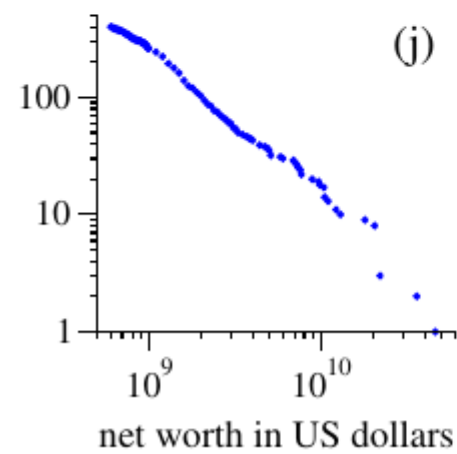
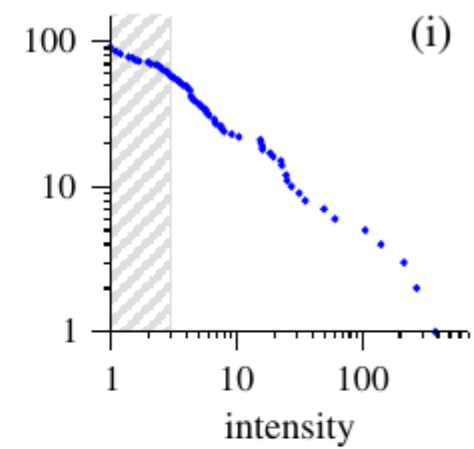
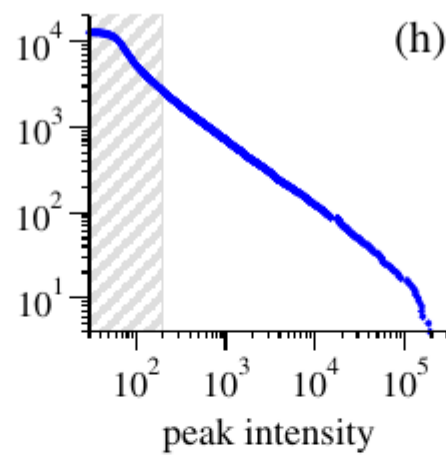
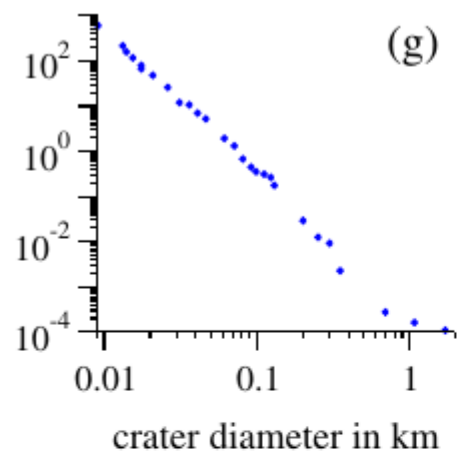
różni autorzy:



A. Orczyk, praca mgr (2008)



M.E.J. Newman, Contemporary Physics **46**, 323 (2005)



M.E.J. Newman, Contemporary Physics **46**, 323 (2005)

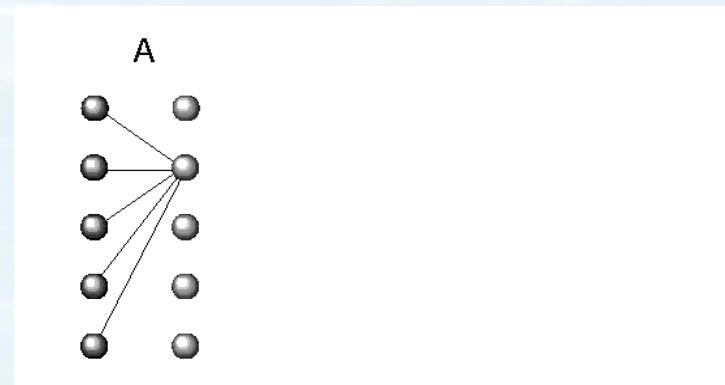
OPTIMALIZACJA TRANSFERU INFORMACJI

Zasada najmniejszego wysiłku:

G.K. Zipf, „*Human Behavior and the Principle of Least Effort*” (1949)

Do komunikacji potrzebne są dwa rodzaje wysiłku:

- wysiłek mówcy, dążącego do przekazania informacji przy użyciu jak najmniejszej liczby słów
- wysiłek słuchacza, dążącego do odbioru jak najprecyzyjniejszego komunikatu (jak największej liczby słów)



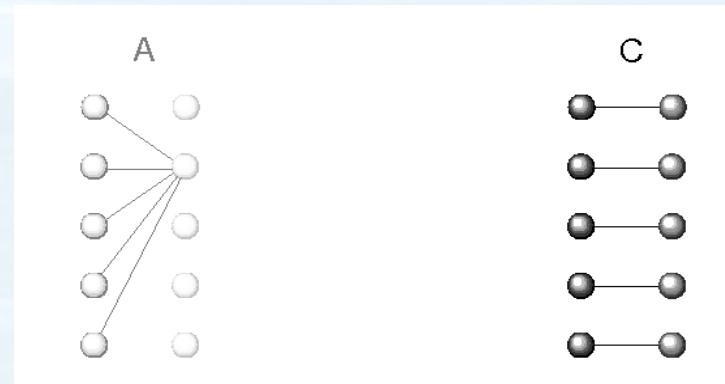
OPTYMALIZACJA TRANSFERU INFORMACJI

Zasada najmniejszego wysiłku:

G.K. Zipf, „*Human Behavior and the Principle of Least Effort*” (1949)

Do komunikacji potrzebne są dwa rodzaje wysiłku:

- wysiłek mówcy, dążącego do przekazania informacji przy użyciu jak najmniejszej liczby słów
- wysiłek słuchacza, dążącego do odbioru jak najprecyzyjniejszego komunikatu (jak największej liczby słów)



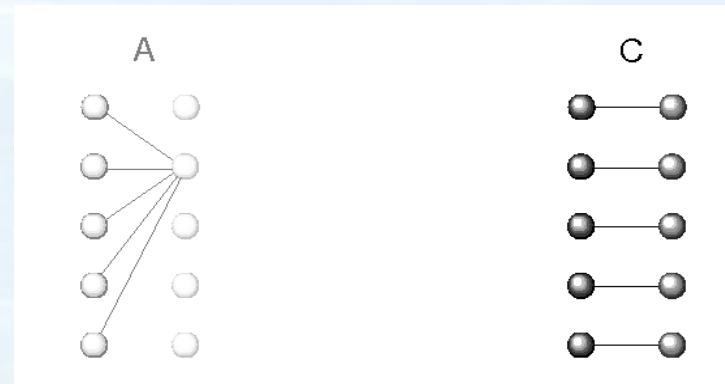
OPTYMALIZACJA TRANSFERU INFORMACJI

Zasada najmniejszego wysiłku:

G.K. Zipf, „*Human Behavior and the Principle of Least Effort*” (1949)

Do komunikacji potrzebne są dwa rodzaje wysiłku:

- wysiłek mówcy, dążącego do przekazania informacji przy użyciu jak najmniejszej liczby słów
- wysiłek słuchacza, dążącego do odbioru jak najprecyzyjniejszego komunikatu (jak największej liczby słów)



OPTYMALIZACJA TRANSFERU INFORMACJI

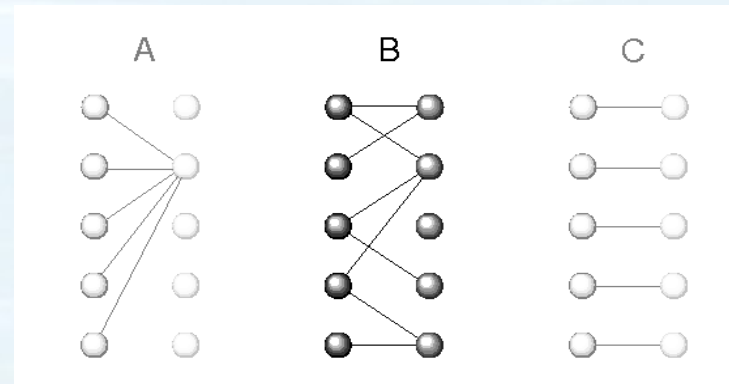
Zasada najmniejszego wysiłku:

G.K. Zipf, „*Human Behavior and the Principle of Least Effort*” (1949)

Do komunikacji potrzebne są dwa rodzaje wysiłku:

- wysiłek mówcy, dążącego do przekazania informacji przy użyciu jak najmniejszej liczby słów
- wysiłek słuchacza, dążącego do odbioru jak najprecyzyjniejszego komunikatu (jak największej liczby słów)

Te dążenia są sprzeczne ze sobą!



OPTIMALIZACJA TRANSFERU INFORMACJI

Zasada najmniejszego wysiłku:

G.K. Zipf, „*Human Behavior and the Principle of Least Effort*” (1949)

Do komunikacji potrzebne są dwa rodzaje wysiłku:

- wysiłek mówcy, dążącego do przekazania informacji przy użyciu jak najmniejszej liczby słów
- wysiłek słuchacza, dążącego do odbioru jak najprecyzyjniejszego komunikatu (jak największej liczby słów)

Te dążenia są sprzeczne ze sobą!

Prawo Zipfa pojawia się jako kompromis pomiędzy mówcą a słuchaczem.

JAK WYTŁUMACZYĆ PRAWO ZIPFA?

Model: Optymalizacja przekazu informacji

B. Mandelbrot, Word **10**, 27 (1954)

p_R	→	prawdopodob. użycia słowa o randze R
C_R	→	koszt użycia słowa R
d_R	→	długość słowa R

Założenie: każdy znak jest jednakowo prawdopodobny: $C_R \sim \log_d R$

średni koszt na słowo: $C = \sum_R p_R C_R$

średnia informacja na słowo: $H = - \sum_R p_R \ln p_R$

minimalizacja $A=C/H$ ze wzgl. na p_R

→ Rezultat: $p_R \sim R^{-\alpha}$

Korpus 65 książek w jęz. polskim:

	Wszystkie		Czasowniki		Rzeczowniki		Przymiotniki	
R	f	Słowo	f	Słowo	f	Słowo	f	Słowo
1	115375	i	19601	jest	15501	harry	1771	cały
2	106408	się	9458	był	10083	pan	1371	polskiej
3	93205	nie	8986	było	7874	pani	1355	pierwszy
4	92958	w	8048	może	4481	oczy	1344	drugi
5	78183	na	7676	ma	4449	ron	1194	innych
6	70989	z	7449	powiedział	4064	potem	1044	stary
7	48998	to	5232	była	3992	kto	1037	całą
8	46456	do	4932	będzie	3747	panie	961	wielki
9	45392	że	4703	miał	3480	hermiona	907	całe
10	43533	a	4557	być	3384	raz	861	inne
11	26408	o	4328	są	3265	geralt	833	wielkie
12	25794	jak	3982	jestem	3194	wokulski	814	dobra
13	24875	co	3802	mam	3133	chwili	754	innego
14	24344	ale	3679	wiem	2918	dumbledore	745	dobry
15	20972	po	3659	rzekł	2726	pana	735	inni
16	20962	za	2956	powiedziała	2708	drzwi	730	cała
17	19601	jest	2550	mógł	2661	ego	717	inny
18	19148	tak	2299	były	2614	jeden	698	pewien
19	15501	harry	2240	odparł	2603	czas	672	młody
20	14718	od	2220	jesteś	2426	panna	671	wirtualnej
21	14370	go	2193	zapytał	2412	głowę	640	innym
22	13660	już	2159	mówił	2375	domu	619	źle
23	13024	ze	2154	wiesz	2359	reynevan	582	żaden
24	12417	mnie	2116	wie	2301	ludzi	581	żadnych
25	11943	jej	2098	masz	2280	czasu	572	pewnego
26	11888	mu	2022	miała	2229	nikt	552	całej
27	11693	tym	1987	proszę	2198	twarz	539	wielką
28	11628	mi	1872	zaczął	2168	głos	531	żadnego
29	11484	jego	1843	chciał	2086	prawda	526	pierwsze
30	11024	ja	1774	mówi	2012	sposób	521	dobrze
31	10898	tylko	1759	mogę	1969	człowiek	520	czarne
32	10427	czy	1711	wiedział	1956	lat	520	jasne
33	10083	pan	1688	chce	1899	ręce	517	stała
34	9851	przez	1633	będę	1875	ludzie	512	wielkim
35	9458	był	1618	musi	1830	rękę	509	czarny
36	9287	tego	1596	spojrzał	1817	chwilę	497	wielkiej
37	8986	było	1494	odpowiedział	1806	głową	496	mały
38	8672	bo	1490	chcę	1757	tyle	473	pewny
39	8542	sobie	1442	byli	1738	dzień	471	drugim
40	8421	gdy	1409	myśli	1706	rzeczy	470	ostatni
41	8150	by	1295	powiedzieć	1687	prawie	470	całym

JAK WYTŁUMACZYĆ PRAWO ZIPFA?

Model: Proces Yule'a-Simona (preferencyjne przyłączanie)

G.U. Yule, Philos. Trans. R. Soc. London B **213**, 21 (1925)

H. Simon, Biometrika **42**, 425 (1955)

Założenia:

- w każdej chwili czasu do tekstu o długości n dodane może być nowe słowo (z prawdopodob. α) lub jedno z dotychczasowych słów (z prawdopodob. $1-\alpha$)
- prawdopodob. dodania dotychczasowego słowa R jest równe: $k_R / \sum_R k_R$ gdzie k_R to liczba powtórzeń słowa R .

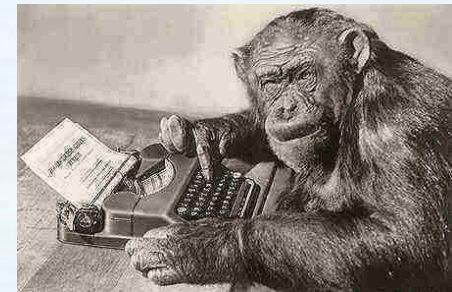
dla $n \rightarrow \infty$ i dla dużych k_R : $p_k \sim k^{-\alpha}$, $\alpha = 2$

(Odwrotne prawo Zipfa)

JAK WYTŁUMACZYĆ PRAWO ZIPFA?

Model: Proces przejściowej ciszy

G.A. Miller, Amer. J. Psychol. **70**, 311 (1957)



Założenia:

- liczba znaków w alfabecie: d
- prawdopodob. wybrania spacji: q
- prawdopodob. wybrania innego znaku niż spacja: $(1-q)/d$

$p_c = q [(1-q)/d]^c$ – prawdopodob. napisania słowa o długości c

słowo o rzędzie R_c ma prawdopodob.:

$$p_c = q \left(\frac{1-q}{d} \right)^c = q \left(\frac{1-q}{d} \right)^{\log_d R_c} = q R_c^{\log_c (1-q) - 1} = q R_c^{-\alpha}$$

JAK WYTŁUMACZYĆ PRAWO ZIPFA?

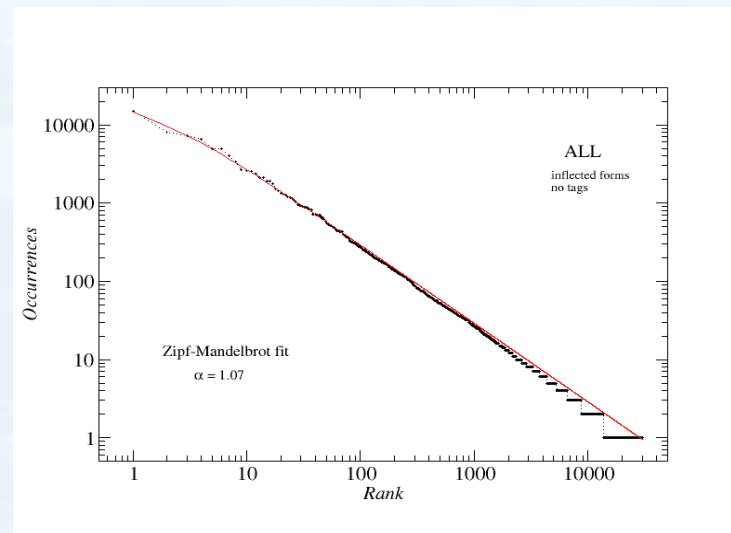
Uogólnienie prawa Zipfa: **prawo Zipfa-Mandelbrota**

B. Mandelbrot, "*Information Theory and Psycholinguistics*",
in: *Scientific Psychology* (Basic Books, 1965)

→ Opisuje odchylenie od skalowania dla niskich rang.

→ Jest rozwiązaniem równania: $\frac{dF(R)}{dR} = -a[F(R)]^\mu, \quad \mu = 1 + 1/\alpha$

→
$$F(R) = \frac{A}{(R + C)^\alpha}$$



JAK WYTŁUMACZYĆ PRAWO ZIPFA?

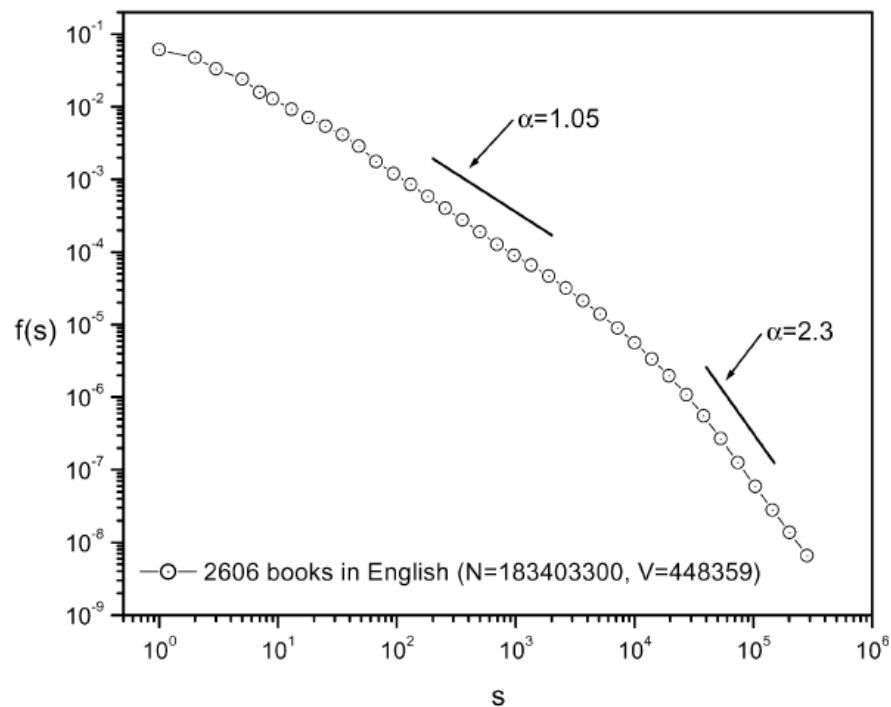
Uogólnienie prawa Zipfa: **dwa obszary skalowania**

M.M. Montemurro, Physica A **300**, 567 (2001)

→ Opisuje zachowanie rozkładu $F(R)$ dla dużych rang

→ Jest rozwiązaniem równania:
$$\frac{dF(R)}{dR} = b[F(R)]^\nu - (a - b)[F(R)]^\mu$$

→
$$\begin{aligned} F(R) &\sim [(\mu - 1)aR]^{1/(1-\mu)}, & \text{dla } F(R) \ll (a/b - 1)^{1/(\mu-\nu)} \\ F(R) &\sim [(\nu - 1)aR]^{1/(1-\nu)}, & \text{dla } F(R) \gg (a/b - 1)^{1/(\mu-\nu)} \end{aligned}$$



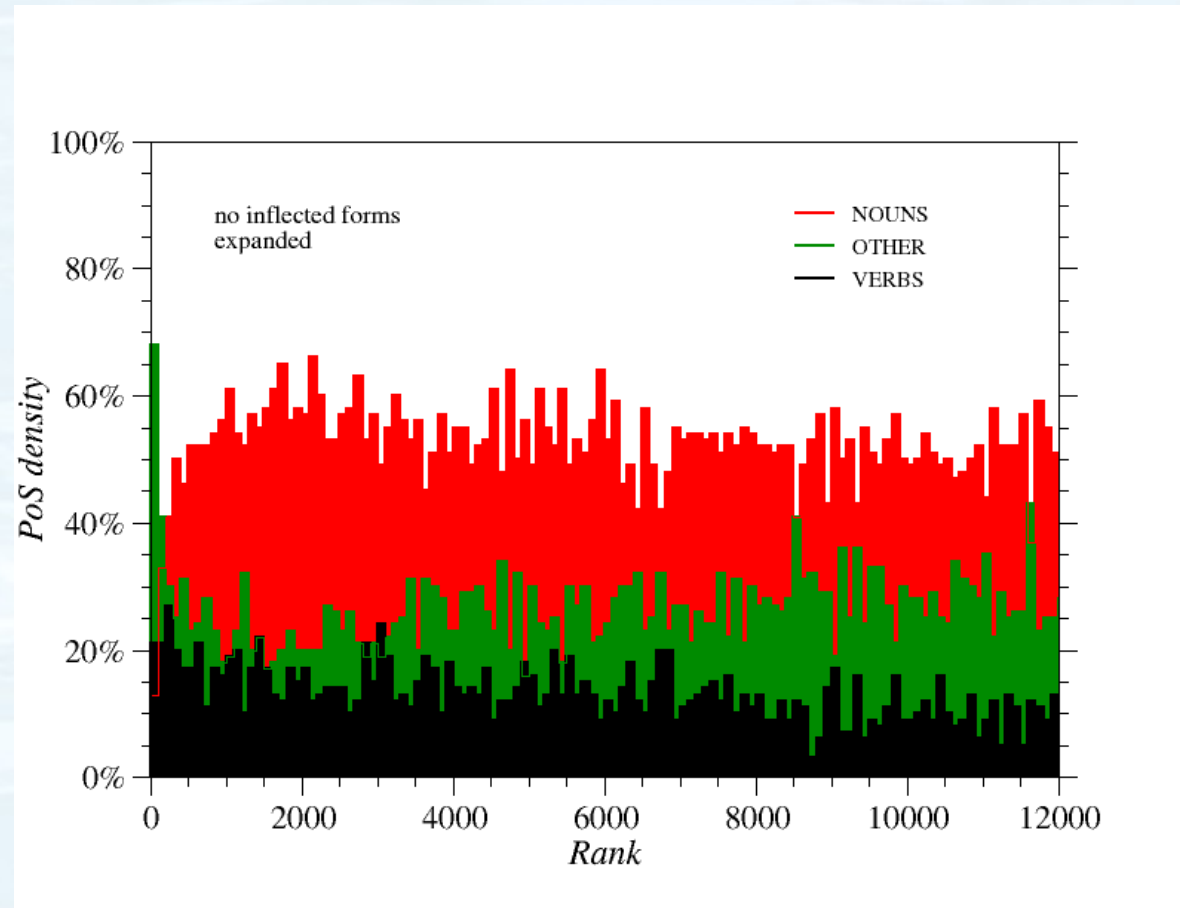
2606 książek w jęz. angielskim

- leksykon podstawowy (wspólny)
- leksykon specjalistyczny

Korpus 65 książek w jęz. polskim:

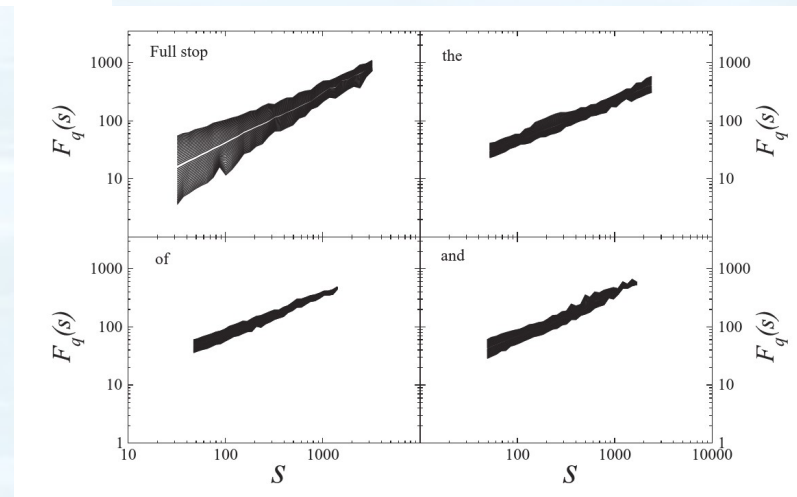
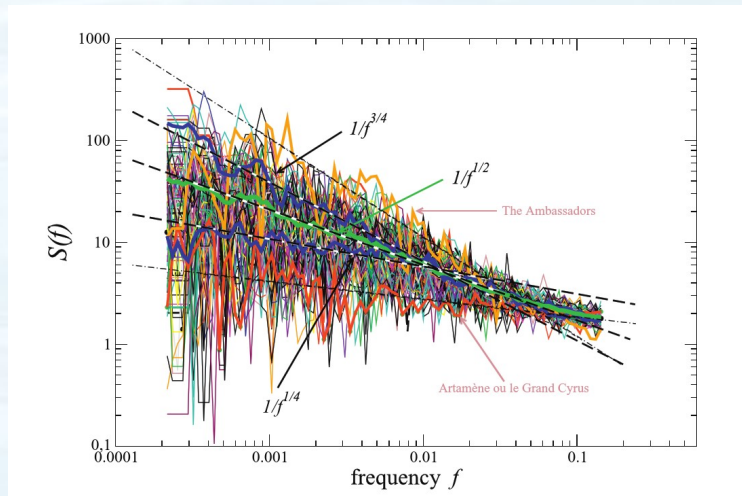
	Wszystkie		Czasowniki		Rzeczowniki		Przymiotniki	
R	f	Słowo	f	Słowo	f	Słowo	f	Słowo
1	115375	i	19601	jest	15501	harry	1771	cały
2	106408	się	9458	był	10083	pan	1371	polskiej
3	93205	nie	8986	było	7874	pani	1355	pierwszy
4	92958	w	8048	może	4481	oczy	1344	drugi
5	78183	na	7676	ma	4449	ron	1194	innych
6	70989	z	7449	powiedział	4064	potem	1044	stary
7	48998	to	5232	była	3992	kto	1037	całą
8	46456	do	4932	będzie	3747	panie	961	wielki
9	45392	że	4703	miał	3480	hermiona	907	całe
10	43533	a	4557	być	3384	raz	861	inne
11	26408	o	4328	są	3265	geralt	833	wielkie
12	25794	jak	3982	jestem	3194	wokulski	814	dobra
13	24875	co	3802	mam	3133	chwili	754	innego
14	24344	ale	3679	wiem	2918	dumbledore	745	dobry
15	20972	po	3659	rzekł	2726	pana	735	inni
16	20962	za	2956	powiedziała	2708	drzwi	730	cała
17	19601	jest	2550	mógł	2661	ego	717	inny
18	19148	tak	2299	były	2614	jeden	698	pewien
19	15501	harry	2240	odparł	2603	czas	672	młody
20	14718	od	2220	jesteś	2426	panna	671	wirtualnej
21	14370	go	2193	zapytał	2412	głowę	640	innym
22	13660	już	2159	mówił	2375	domu	619	złe
23	13024	ze	2154	wiesz	2359	reynevan	582	żaden
24	12417	mnie	2116	wie	2301	ludzi	581	żadnych
25	11943	jej	2098	masz	2280	czasu	572	pewnego
26	11888	mu	2022	miała	2229	nikt	552	całej
27	11693	tym	1987	proszę	2198	twarz	539	wielką
28	11628	mi	1872	zaczął	2168	głos	531	żadnego
29	11484	jego	1843	chciał	2086	prawda	526	pierwsze
30	11024	ja	1774	mówi	2012	sposób	521	dobrze
31	10898	tylko	1759	mogę	1969	człowiek	520	czarne
32	10427	czy	1711	wiedział	1956	lat	520	jasne
33	10083	pan	1688	chce	1899	ręce	517	stała
34	9851	przez	1633	będę	1875	ludzie	512	wielkim
35	9458	był	1618	musi	1830	rękę	509	czarny
36	9287	tego	1596	spojrzał	1817	chwilę	497	wielkiej
37	8986	było	1494	odpowiedział	1806	głową	496	mały
38	8672	bo	1490	chcę	1757	tyle	473	pewny
39	8542	sobie	1442	byli	1738	dzień	471	drugim
40	8421	gdy	1409	myśli	1706	rzeczy	470	ostatni
41	8150	by	1295	powiedzieć	1687	prawie	470	całym

James Joyce, *Ulysses*



Rola znaków interpunkcyjnych w tekście:

- ▶ pozwalają wyróżnić logiczne bloki w ramach większej całości
- ▶ usuwają niejednoznaczności w treści
- ▶ wskazują miejsca, w których można zaczerpnąć oddech
- ▶ sugerują intonację i melodię wypowiedzi
- ▶ pozwalają na zakodowanie emocji (np. ironia, zdziwienie)
- ▶ określają szybkość biegu zdarzeń
- ▶ podkreślają prostotę lub złożoność sytuacji
- ▶ mogą określać pewne aspekty stylu
- ▶ są jednym z nośników korelacji dalekozasięgowych w tekście

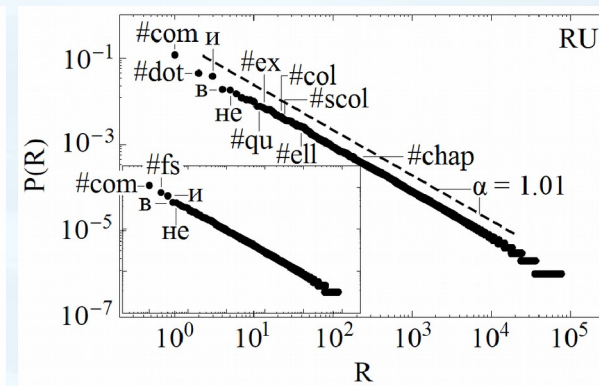
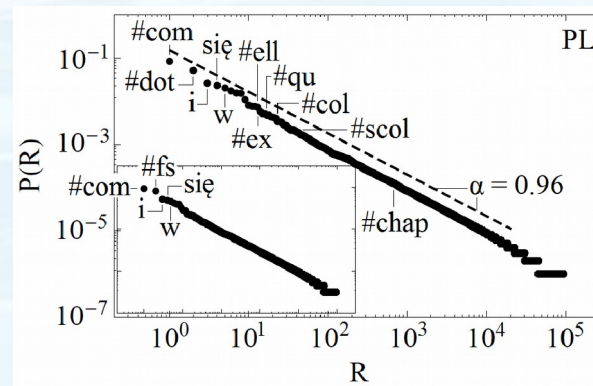
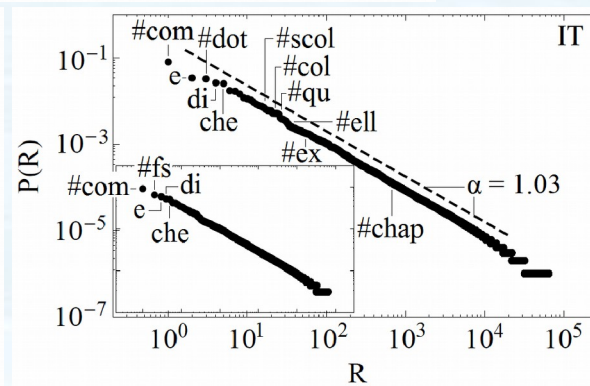
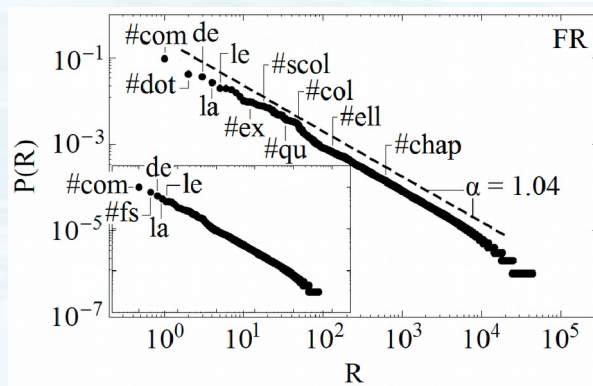
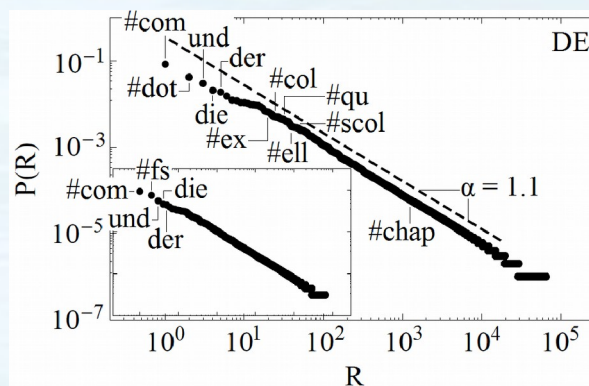
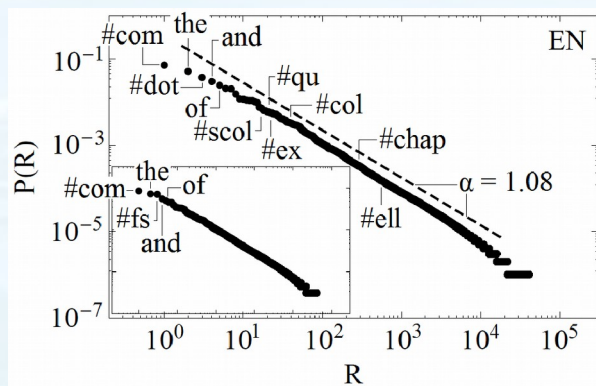


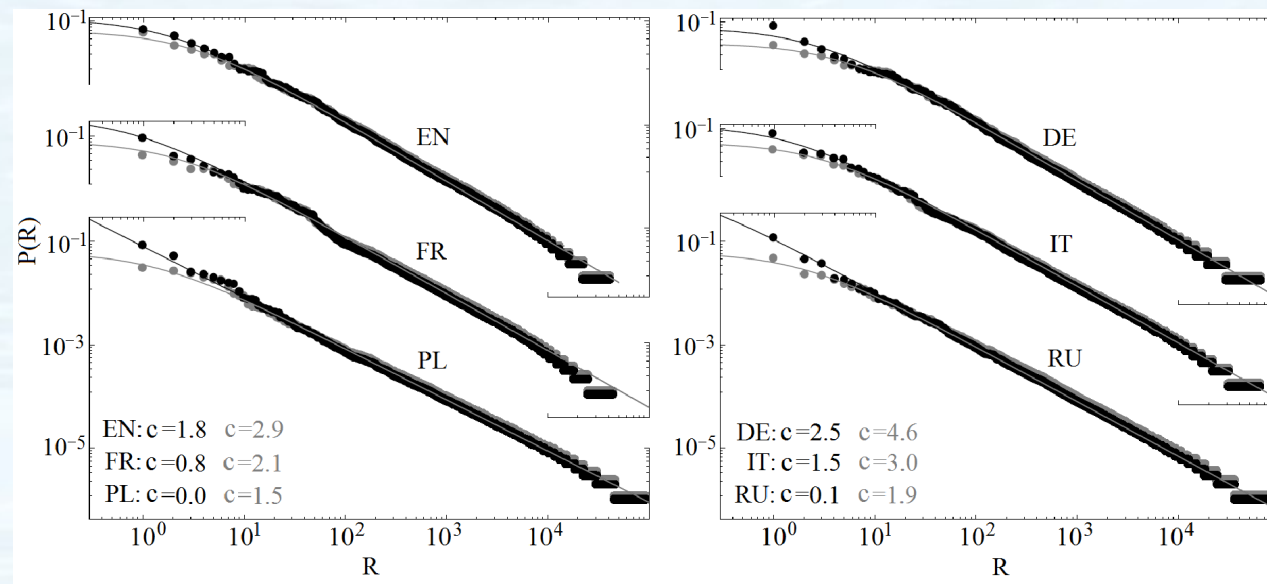
S. Drożdż *et al.*, Information Sciences **331**, 32-44 (2016)

Skoro znaki interpunkcyjne niosą w sobie zawartość znaczeniową, to czy odwrotna sytuacja jest możliwa?



Wniosek: potraktujmy znaki interpunkcyjne, jak gdyby były zwykłymi słowami.





A. Kulig *et al.*, Information Sciences **375**, 98-113 (2017)

Sieci sąsiedztwa słów:

unikalne słowo



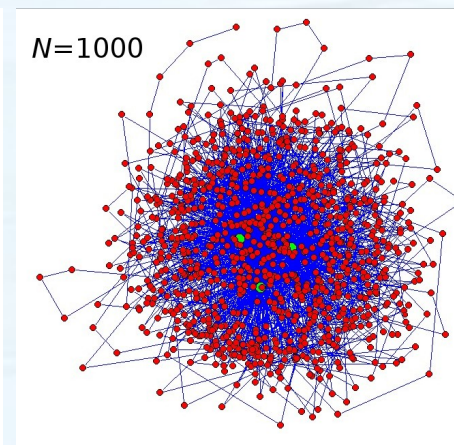
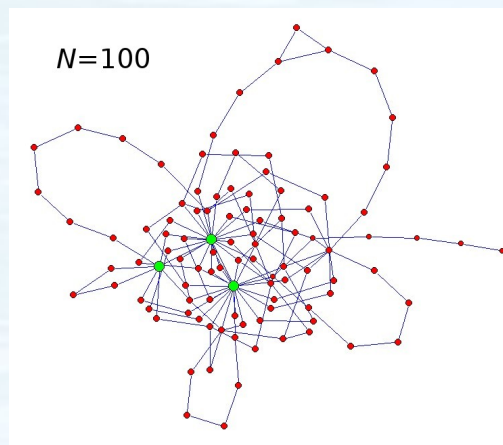
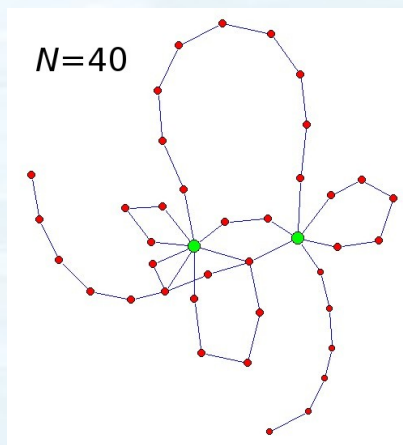
węzeł sieci

sąsiedztwo dwóch słów co najmniej raz w tekście

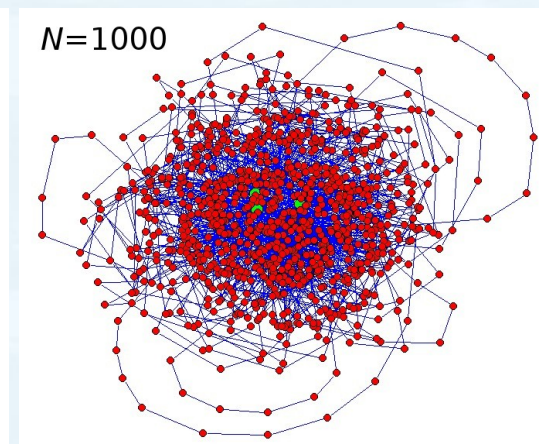
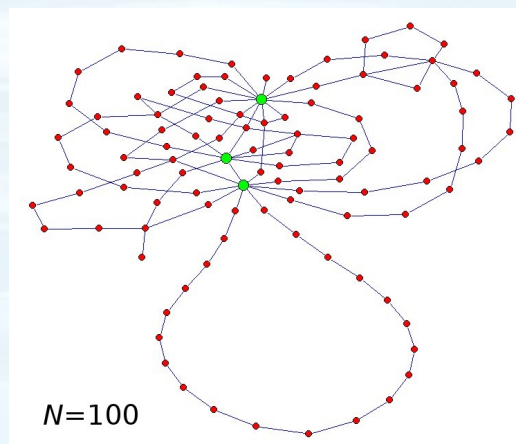
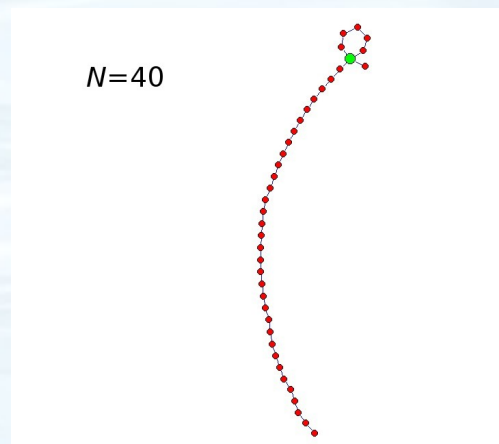


krawędź

Język angielski (James Joyce, "Ulysses"):



Język polski (Bolesław Prus, "Lalka"):



Sieci sąsiedztwa słów:

unikalne słowo	→	węzeł sieci
sąsiedztwo dwóch słów co najmniej raz w tekście	→	krawędź

Opis lokalnej topologii sieci:

średnia najkrótsza ścieżka dla węzła i :

$$\ell_i = \frac{1}{N-1} \sum_{j=1, j \neq i}^N d(i, j)$$

lokalny współczynnik gronowania dla węzła i :

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

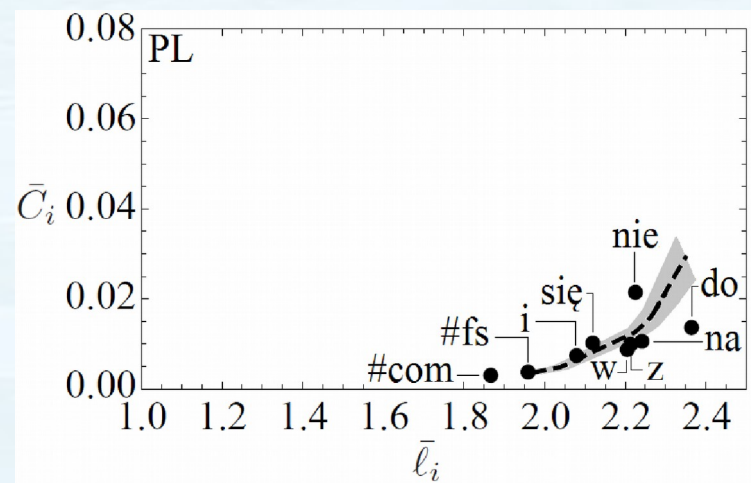
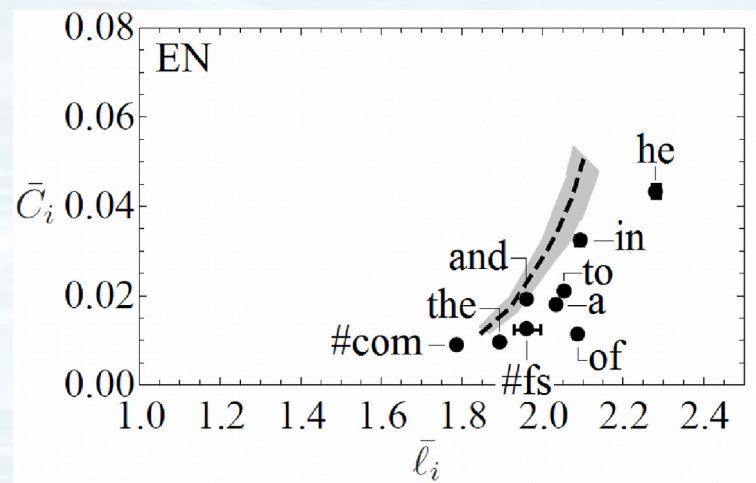
e_i - liczba krawędzi między sąsiadami węzła i

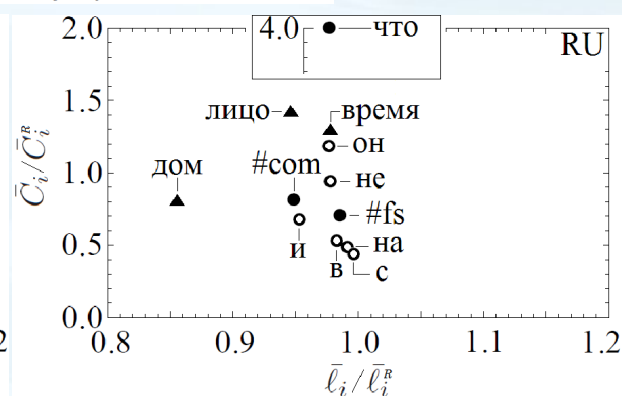
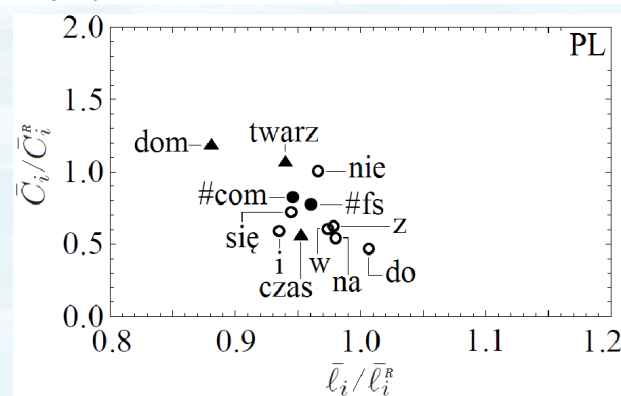
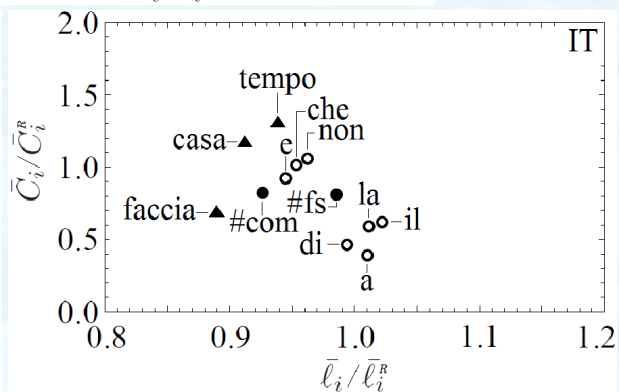
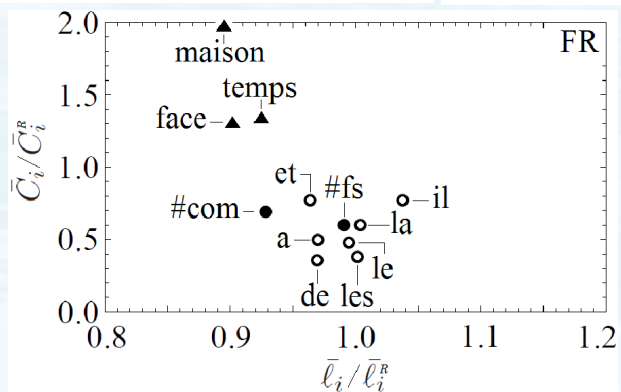
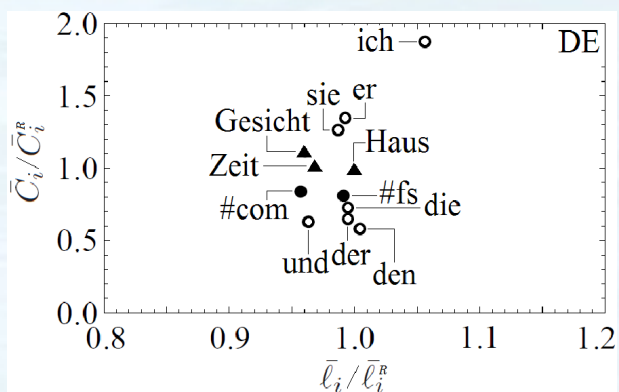
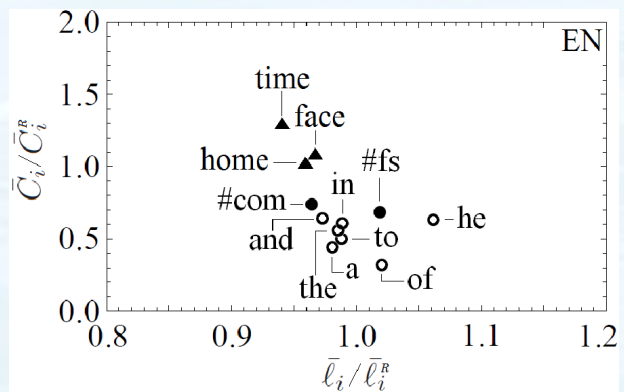
k_i - krotność węzła i

$d(i, j)$ - odległość między węzłami i, j

$\ell_i(k_i)$ $\longrightarrow$ funkcja malejąca

$$C_i(k_i) \sim k_i^{-\gamma}, \quad \gamma > 0$$





Podsumowanie:

- rozkład częstości występowania słów w zależności od rangi opisuje potęgowe prawo Zipfa;
- poprawkę dla niskich rang (tj. dla najczęstszych słów) opisuje prawo Zipfa-Mandelbrota;
- różne języki charakteryzują się różnymi odchyleniami od zależności potęgowej $F(R)$ dla niskich rang;
- uwzględnienie znaków interpunkcyjnych pozwala na przywrócenie skalowania w zakresie niskich rang;
- występują pewne różnice między językami, przy czym język angielski wykazuje najmniejszą zgodność z zachowaniem potęgowym;
- znaki interpunkcyjne posiadają podobne własności statystyczne do własności regularnych słów;
- znaki interpunkcyjne powinno się uwzględniać w statystycznej analizie języka pisanego na równi ze zwykłymi słowami.