Prawo Zipfa
próby objaśnień

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

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.

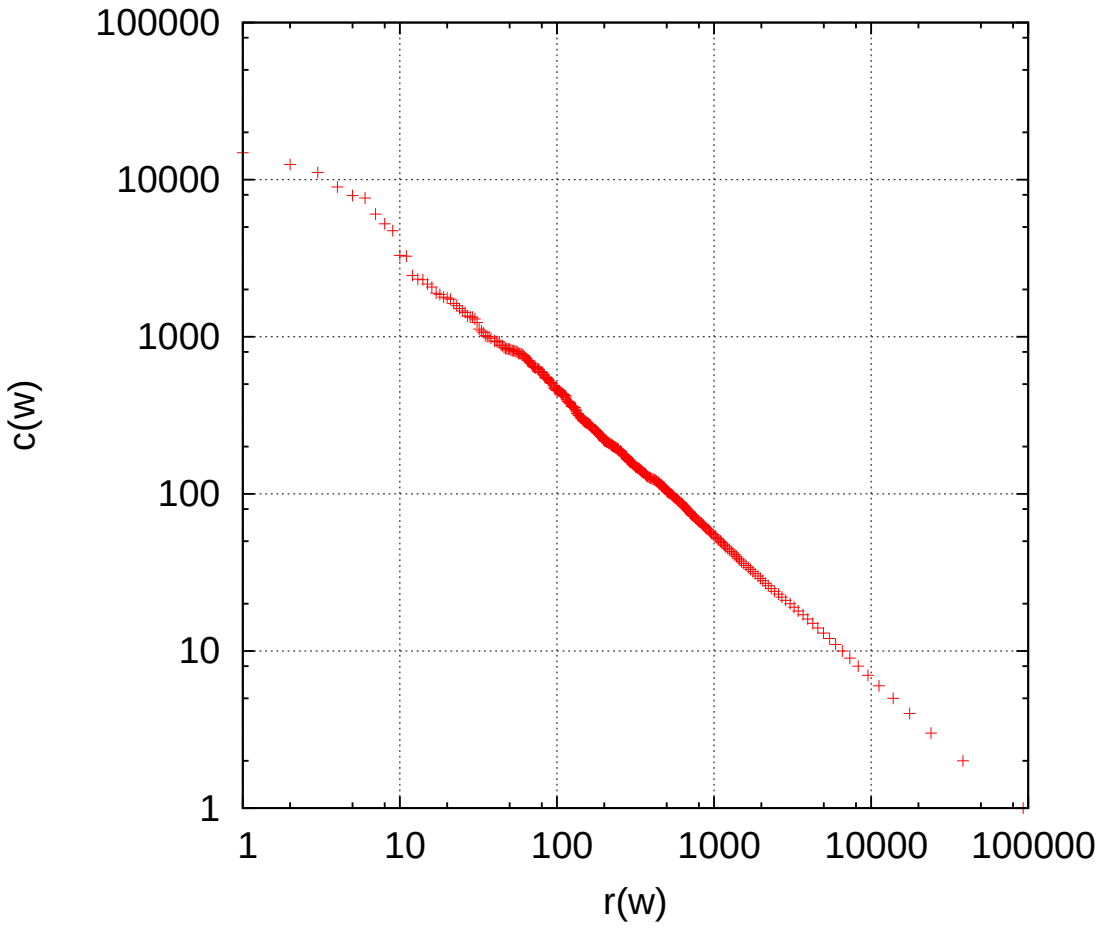
Ciąg dalszy listy rangowej

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,

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

Najsłynniejsze prawo lingwistyki kwantytatywnej

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

Wykresy dla listy rangowej liter wyglądają inaczej!

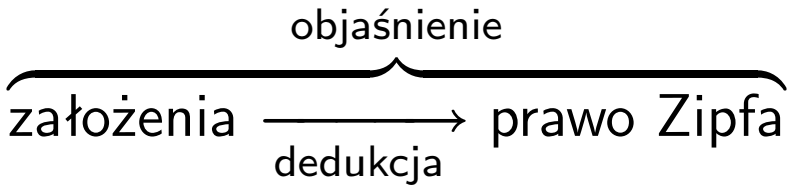
Rozkłady ranga-częstość spełniające prawo Zipfa pojawiają się także poza lingwistyką:

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

Potrzeba objaśnień

Dlaczego prawo Zipfa obowiązuje
dla tak wielu rozkładów ranga-częstość?

Czym są objaśnienia?



Założenia mogą być różnorakiej natury:

- istnieje pewien probabilistyczny model tekstu,
- tekst jest efektem optymalizacji pewnych kryteriów,
- słowa w tekście można wyodrębnić za pomocą pewnego algorytmu.

Niektóre założenia prowadzące do prawa Zipfa mogą być fałszywe.
Może istnieć wiele niezależnych prawdziwych założeń.

Stan badań z lotu ptaka

Odkryto kilka rozumowań dedukcyjnych, które objaśniają pojawianie się prawa Zipfa w rozkładach ranga-częstość.

Założenia czynione w klasycznych objaśnieniach są:

- bądź to fałszywe (w odniesieniu do ludzkich tekstów),
- bądź to tajemnicze (same wymagają weryfikacji).

Dotychczasowe objaśnienia to "inspirujące potwory".

Potrzeba dalszych badań na pograniczu lingwistyki i matematyki (probabilistyka, teoria informacji, informatyka teoretyczna).

Cztery objaśnienia

Omówię 4 modele, w których pojawia się prawo Zipfa:

- model „małpy przy klawiaturze” (Mandelbrot (?), Miller),
- losowanie ze sprzężeniem zwrotnym (Simon),
- minimalizacja średniej długości na bit informacji (Mandelbrot),
- model najmniejszej gramatyki (Dębowski, praca w toku).

Caveat emptor

Czy nie zamierzamy objaśnić czegoś, co nie istnieje?

- Niektóre pozalingwistyczne przykłady obciętych rozkładów ranga-częstość zebrane przez Zipfa (1949) są znacznie lepiej modelowane przez modele inne niż prawo Zipfa (np. przez rozkład log-normalny).
- W odniesieniu do listy rangowej słów, prawo Zipfa jest znacznie lepszym modelem niż rozkład log-normalny.

cytowane za:

- Richard Perline (2005), *Strong, Weak and False Inverse Power Laws*, Statistical Science 20:68–88.

George A. Miller (1920-)

Literatura

- Benoit B. Mandelbrot (1953), *An informational theory of the statistical structure of languages*, [w:] W. Jackson (red.), Communication Theory (486–502). (?)
- George A. Miller (1957), *Some effects of intermittent silence*, American Journal of Psychology 70:311–314.
- George A. Miller (1965), *Introduction*, [w:] George Kingsley Zipf, *Human Behavior and the Principle of Least Effort* (II wydanie).

Model „małpy przy klawiaturze”

Założenia:

- Tekst jest ciągiem liter i odstępów.
- Tekst otrzymujemy wciskając losowo klawisze klawiatury.
- Słowa w tekście to ciągi liter od odstępu do odstępu.

Rezultat:

- Lista rangowa słów spełnia prawo Zipfa-Mandelbrota.

Zgrubne uzasadnienie

N — długość tekstu w znakach
L + 1 — liczba różnych znaków
Prawdopodobieństwo konkretnego znaku: $\mathbf{1/(L+1)}$.
Oczekiwana liczba słów (= odstępów) w tekście: $\mathbf{N/(L+1)}$.

Niech **w** będzie **n**-literowym słowem o najwyższej randze.

Oszacowanie jego częstości **c(w)** i rangi **r(w)**:

$$\mathbf{c(w)} \approx \frac{1}{(L+1)^n} \cdot \frac{1}{L+1} \cdot \frac{N}{L+1},$$

$$\mathbf{r(w)} \approx 1 + L + L^2 + \ldots L^n = \frac{L^{n+1}-1}{L-1} \approx \frac{L^{n+1}}{L-1}.$$

$$\mathbf{c(w)} \approx \frac{N}{L+1} \cdot \left[\frac{1}{(L-1)\cdot r(w)}\right]^{\log(L+1)/\log L}$$

Niedostatki modelu „małpy przy klawiaturze” (I)

I. Wszystkie słowa długości n mają podobną częstość, tzn.

- na wykresie ranga-częstość widać poziome pasy.

Pochyloną prostą widać dla znaków **o nierównych p-stwach**:

- T.C. Bell, J.G. Cleary, I.H. Witten (1990), *Text Compression*.
- Brian Conrad, Michael Mitzenmacher (2004), *Power laws for monkeys typing randomly: The case of unequal probabilities*, IEEE Transactions on Information Theory 50:1403–1414.

Niedostatki modelu „małpy przy klawiaturze” (II)

II. Proces stochastyczny w tym modelu jest ergodyczny, tzn.

- ustalone słowo ma bardzo podobną częstość w dowolnym dostatecznie dużym tekście (korpusie).

Jest to sprzeczne z tym, co wiemy o języku naturalnym:

- zróżnicowanie słownikowe tekstów jest **banalnym faktem**.

Możliwe jest rozróżnianie autorów tekstów na podstawie częstości samych spójników i przyimków:

- Adam Pawłowski (2003), *O problemie atrybucji tekstów w lingwistyce kwantytatywnej*, [w:] J. Linde-Usiekniewicz, R. Huszcza, *Prace językoznawcze dedykowane Profesor Jadwidze Sambor*. (169–190)

Herbert A. Simon (1916–2001)

Otrzymał nagrodę Nobla z ekonomii w 1978 roku.

Literatura

- G. Udny Yule (1924), *A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis F.R.S*. Philosophical Transactions of the Royal Society, London Ser. B 213:21–87.
- Robert Gibrat (1931), *Les inégalités économiques*.
- Herbert A. Simon (1955), *On a class of skew distribution functions*, Biometrika, 42:425–440.

Losowanie ze sprzężeniem zwrotnym

Założenia:

- Tekst jest ciągiem słów.
- Tekst otrzymujemy losując kolejne słowa.
- Gdy losujemy słowo na **n**-tej pozycji tekstu,
  - p-stwo wylosowania zaświadczonego już słowa jest proporcjonalne do jego częstości na pozycjach od **1** do $\mathbf{(n-1)}$ (**prawo Gibrata**);
  - p-stwo wylosowania nowego słowa wynosi $\alpha > \mathbf{0}$.

Rezultat:

- Rozkład częstości-częstości słów ma rozkład Yule'a (uogólnienie prawa Lotki).

Prawo Lotki (równoważne prawu Zipfa)

A — liczba różnych słów (typów) w tekście
N — liczba wszystkich słów (okazów)
F(c) — liczba typów słów o częstości **c**
c(r) — częstość słowa o randze **r**
r(c) — największa ranga słowa o częstości **c**

Z prawa Zipfa

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

mamy $\mathbf{r(c)} = \mathbf{A/c}$ i prawo Lotki (zwane też rozkładem Simona)

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

Rozkład Yule'a

W granicy nieskończonego tekstu losowanie ze sprzężeniem zwrotnym daje rozkład Yule'a:

$$\lim_{N\to\infty} \frac{F(c)}{A} = \frac{1}{1-\alpha} \cdot \frac{1 \cdot 2 \cdot \ldots \cdot (c-1)}{(1+\rho) \cdot (2+\rho) \cdot \ldots \cdot (c+\rho)},$$

gdzie $\rho = 1/(1-\alpha)$ oraz $\lim_{N\to\infty} A/N = \alpha$.

Dla $\alpha \to 0$ rozkład Yule'a sprowadza się do prawa Lotki:

$$\lim_{N\to\infty} \frac{F(c)}{A} \xrightarrow[\alpha\to 0]{} \frac{1}{c(c+1)}.$$

Niedostatki losowania ze sprzężeniem zwrotnym

Słowa w tekście mają tendencję do powtarzania się.

Wydawać by się mogło, że losowanie ze sprzężeniem zwrotnym lepiej uwzględnia tę tendencję niż model „małpy przy klawiaturze".

W rzeczywistości jest odwrotnie.

	$F(1)/A$	$\lim_N A/N$
idealne prawo Zipfa (Lotki)	$= 1/2$	$= 0$
model „małpy przy klawiaturze"	$< 1/2$	$= 0$
losowanie ze sprzężeniem zwrotnym	$> 1/2$	> 0

Efekt ten jest skutkiem stałego p-stwa wylosowania nowego słowa.

Prawo Gibrata z mniejszą liczbą hapaksów

Rozpatrzmy modyfikację losowania ze sprzężeniem zwrotnym:

- p-stwo wylosowania nowego słowa na **n**-tej pozycji w tekście wynosi nie α lecz $\alpha_0 \mathbf{n}^{\mathbf{C}-1}$.

Wówczas dla dużych **N** otrzymamy $\mathbf{A} \propto \mathbf{N}^{\mathbf{C}}$.
Rozkład ranga-częstość także jest zbliżony do rzeczywistego.

- Damián H. Zanette, Marcelo A. Montemurro (2005), *Dynamics of text generation with realistic Zipf's distribution*, Journal of Quantitative Linguistics 12:29–40.

Benoit B. Mandelbrot (1924-)

Literatura

- Benoit B. Mandelbrot (1954), *Structure formelle des textes et communication*, Word 10:1–27.
- V. K. Balasubrahmanyan, S. Naranan (2005), *Entropy, information, and complexity*, [w:] Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski (red.) *Quantitative Linguistics. An International Handbook* (878–891).

Minimalizacja średniej długości na bit

Założenia:

- Tekst jest efektem optymalizacji.
  (Tekst nie musi być produktem prostego procesu losowego.)
- Przy ustalonej liczbie okazów słów w tekście rozkład częstości słów maksymalizuje iloraz $\mathbf{H}/\mathbf{C}$, gdzie
  - **H** — entropia słowa w tekście,
  - **C** — średnia długość słowa w tekście.
- Długość słowa o randze **r** jest proporcjonalna do $\mathbf{\log r}$.

Rezultat:

- Lista rangowa słów spełnia prawo Zipfa-Mandelbrota.

Przybliżone uzasadnienie (I)

Entropia słowa w tekście to

$$\mathbf{H} = -\sum_{\mathbf{r}=1}^{\infty} \frac{\mathbf{c(r)}}{\mathbf{N}} \log \frac{\mathbf{c(r)}}{\mathbf{N}},$$

a średnia długość słowa w tekście to

$$\mathbf{C} \approx \sum_{\mathbf{r}=1}^{\infty} \frac{\mathbf{c(r)}}{\mathbf{N}} \log \mathbf{r}.$$

Przybliżone uzasadnienie (II)

Suma $\sum_{r=1}^{\infty} \mathbf{c(r)} = \mathbf{N}$ ma być stała, więc $\mathbf{H/C}$ jest największe dla

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{c(r)}} \left[\frac{\mathbf{H}}{\mathbf{C}}\right] - \mathbf{const}\, \frac{\partial}{\partial \mathbf{c(r)}} \left[\sum_{j=1}^{\infty} \mathbf{c(j)} - \mathbf{N}\right]$$
$$= -\frac{\mathbf{[1 + \log c(r)]C + H \log r}}{\mathbf{NC^2}} - \mathbf{const}\,.$$

Zatem $\mathbf{C \log c(r) + H \log r = const = C \log N}$, czyli:

$$\mathbf{c(r)} = \frac{\mathbf{N}}{\mathbf{r^{H/C}}}$$

Uwagi do minimalizacji średniej długości na bit (I)

I. Uzasadnienie nie uwzględnia dyskretności c(r).

- Dla $\mathbf{c(r) = N/r^{H/C}}$ równość $\sum_{\mathbf{r=1}}^{\infty} \mathbf{c(r) = N}$ zachodzi tylko w granicy $\mathbf{H/C} = \infty$.
- W rzeczywistości częstość $\mathbf{c(r)}$ jest liczbą naturalną, przy czym na pewno $\mathbf{c(r) = 0}$ dla $\mathbf{r > N}$.
- W takim przypadku entropia słowa jest ograniczona przez długość tekstu: $\mathbf{H \leq \log N}$.

Być może otrzymamy $\mathbf{c(r) \approx A/r}$, maksymalizując $\mathbf{H/C}$ dla skończonego tekstu, ale nikt tego poprawnie nie udowodnił.

Uwagi do minimalizacji średniej długości na bit (II)

II. Dlaczegóż tekst miałby maksymalizować H/C?

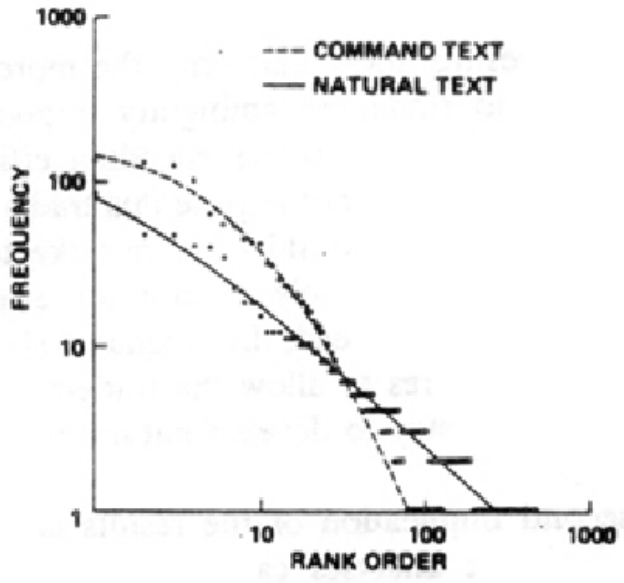
- Czy chodzi o to, aby wszystkie słowa w przeliczeniu na jednostkę długości były równie trudne do odgadnięcia?
- Czy maksymalizacja $\mathbf{H/C}$ zachodzi przy edycji tekstu? (Czy $\mathbf{H_n/C_n < H_{n+1}/C_{n+1}}$ dla kolejnych wersji tekstu?)

- Claude Shannon (1950), *Prediction and entropy of printed English*, Bell System Technical Journal 30:50–64.
- Stephen R. Ellis, Robert J. Hitchcock (1986), *Emergence of Zipf's Law: Spontaneous encoding optimization by users of a command language*, IEEE Transactions on Systems, Man, and Cybernetics SMC-16:423–427.

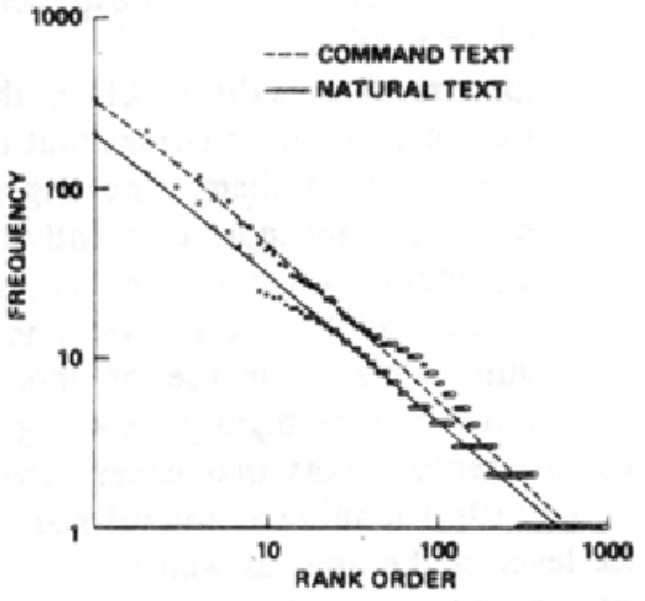
Stephen R. Ellis i Robert J. Hitchcock (1986)

Rozkład częstości poleceń UNIX-a dla dwóch użytkowników:

początkującego



zaawansowanego



Lingwistyka synergetyczna

W kontekście badań ilościowych praw językowych sformułowany był postulat rozwijania podejścia synergetycznego:

> Ilościowe prawa językowe są efektem dążenia systemu językowego do chwiejnej równowagi i są ze sobą powiązane.
>
> *Hipoteza:* Dynamika języka wyjaśnia prawa ilościowe.

- Rolf Hammerl, Jadwiga Sambor (1993), *O statystycznych prawach językowych*.

Czym są słowa tekstowe?

Dotychczasowi „objaśniacze” nie dociekali, czym są słowa.

Słowa to ciągi liter rozdzielone odstępami, prawda?

Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki...

Tylko co zrobić z następującymi tekstami?

ARMAVIRUMQUECANOTROIAEQUIPRIMUSABORIS ITALIAMFATOPROFUGUSLAVINIAQUEVENIT...

人人生而自由，在尊嚴和權利上一律平等。他們賦有理性和良心，並應以兄弟關係的精神互相對待。

W mowie odstępy też są często pomijane.

Szukajmy innej definicji formalnej!

Trudno oczekiwać, aby proste wyjaśnienie dedukcyjne uwzględniało fakt, że słowa są jednostkami definiowanymi przez znaczenie.

Niesatysfakcjonujące jest opieranie wyjaśnienia na fakcie, że słowa są elementami pewnego słownika danego a priori.

Możemy jednak szukać innej definicji słowa jako ciągu znaków:
Podział tekstu na takie ciągi znaków ma być zbliżony
do podziału tekstu na słowa rozumiane tradycyjnie.

Być może dla takiej definicji słowa tekstowego prawo Zipfa

- zachodziłoby empirycznie **lub**
- byłoby matematyczną tautologią **lub**
- byłoby możliwe do wyprowadzenia z innych własności tekstu.

Gramatyki bezkontekstowe generujące jeden tekst

$$\mathbf{G} = \left\{ \begin{array}{l} \mathbf{A_0} \mapsto \mathbf{A_1 A_1 A_3 A_8 A_3 A_2 A_6 A_4} \\ \mathbf{A_1} \mapsto \mathbf{A_5 A_8 A_5 A_2} \\ \mathbf{A_2} \mapsto \mathbf{A_8 A_6 A_7 , A_4} \\ \mathbf{A_3} \mapsto \texttt{Jeszcze\_raz} \\ \mathbf{A_4} \mapsto \mathbf{A_7}\texttt{\_nam.\_} \\ \mathbf{A_5} \mapsto \texttt{Sto\_lat} \\ \mathbf{A_6} \mapsto \texttt{Niech} \\ \mathbf{A_7} \mapsto \texttt{\_żyje} \\ \mathbf{A_8} \mapsto \texttt{!\_} \end{array} \right\}$$

Sto lat! Sto lat! Niech żyje, żyje nam.
Sto lat! Sto lat! Niech żyje, żyje nam.
Jeszcze raz! Jeszcze raz! Niech żyje, żyje nam.
Niech żyje nam.

Zapis segmentacji tekstu w formie takiejże gramatyki

- Tekst $\mathbf{T}$ jest konkatenacją słów $\mathbf{w_i}$ (okazy).
- Każde słowo $\mathbf{W_i}$ (typ) jest konkatenacją sylab $\mathbf{s_{ij}}$ (okazy).
- Każda sylaba $\mathbf{S_j}$ (typ) jest konkatenacją głosek $\mathbf{a_{jk}}$ (okazy).

„Idealna" gramatyka dla tekstu $\mathbf{T}$:

$$\mathbf{G}^{\text{ideal}} = \left\{ \begin{array}{ll} \mathbf{A_0} \mapsto \mathbf{w_1 w_2 \ldots w_{N(T)}} & \\ \mathbf{W_i} \mapsto \mathbf{s_{i1} s_{i2} \ldots s_{iN(W_i)}}, & \mathbf{i = 1, 2, \ldots, V_W} \\ \mathbf{S_j} \mapsto \mathbf{a_{j1} a_{j2} \ldots a_{jN(S_j)}}, & \mathbf{j = 1, 2, \ldots, V_S} \end{array} \right\}$$

Czy istnieje algorytm, który dla dowolnego tekstu konstruuje gramatykę zbliżoną do gramatyki idealnej?

Model najmniejszej gramatyki

Długość gramatyki **G** to suma długości prawych stron jej reguł:

$$|\mathbf{G}| := \sum_{(\mathbf{A_i} \mapsto \mathbf{g_i}) \in \mathbf{G}} |\mathbf{g_i}|\,, \quad \text{gdzie np.} \quad |\mathbf{A_7}\texttt{\_nam.\_}| = \mathbf{7}.$$

Najmniejsza gramatyka dla tekstu to **G** o najmniejszym $|\mathbf{G}|$. Istnieje wiele szybkich algorytmów obliczających przybliżenia tej gramatyki, które nazywam gramatykami lokalnie najmniejszymi.

- Moses Charikar, Eric Lehman, ..., Abhi Shelat (2005), *The Smallest Grammar Problem*, IEEE Transactions on Information Theory, 51:2554–2576.
- John C. Kieffer, Enhui Yang (2000), *Grammar-based codes: A new class of universal lossless source codes*, IEEE Transactions on Information Theory, 46:737–754.

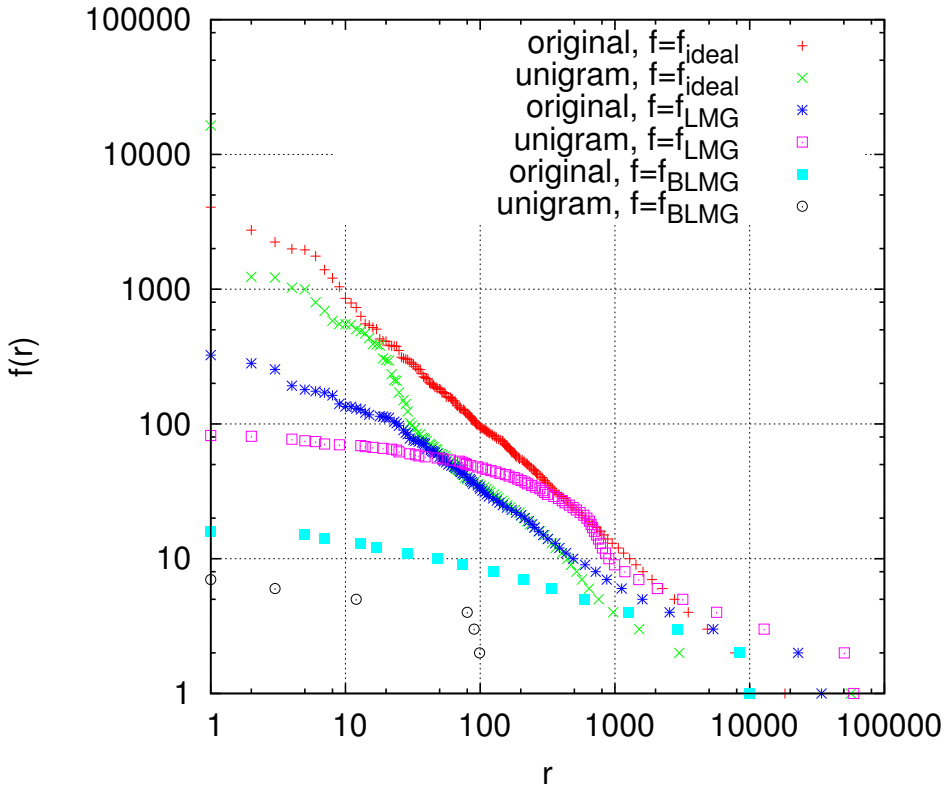
Obserwacja lingwistów komputerowych

Problem najkrótszej gramatyki studiowali od strony praktycznej:

- J. Gerard Wolff (1980), *Language acquisition and the discovery of phrase structure*. Language and Speech, 23:255–269.
- Carl G. de Marcken (1996), *Unsupervised Language Acquisition*.
- Craig G. Nevill-Manning (1996), *Inferring Sequential Structure*.

Badacze ci obliczali pewne lokalnie najmniejsze gramatyki dla korpusów tekstów w j. angielskim i chińskim. Zauważyli, że wiele jednostek nieterminalnych tych gramatyk odpowiada sylabom, morfemom, słowom i ustalonym frazom.

Prawo Zipfa dla gramatyk dla *W pustyni i w puszczy*



Rozkład ranga-częstość silnie zależy od rodzaju gramatyki.
Dla niektórych gramatyk mamy w przybliżeniu prawo Zipfa.
Wykresy dla *Gulliver's Travels* (po angielsku) są b. podobne.

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?

Czy prawa ilościowe dla gramatyk można wyjaśnić?

W pracach:

- Ł. Dębowski (2006), *Menzerath's law for the smallest grammars*, [w:] R. Köhler, P. Grzybek, Viribus Quantitatis,
- Ł. Dębowski (2006), *On Hilberg's law and its links with Guiraud's law*, Journal of Quantitative Linguistics, 13:81–109

próbowałem objaśnić znany wzór empiryczny (prawo Heapsa):

$$\mathbf{A} \propto \mathbf{N}^{\mathbf{C}}, \quad \mathbf{C} \approx \mathbf{1/2},$$

gdzie **A** to liczba typów słów, zaś **N** to liczba okazów słów.

Wzór ten proponowali: W. Kuraszkiewicz, J. Łukaszewicz (1951), P. Guiraud (1954), G. Herdan (1964), H. S. Heaps (1978).

Prawo Heapsa jest konsekwencją prawa Zipfa-Mandelbrota (Kornai 2002; van Leijenhorst, van der Weide 2005).

Dwie nierówności

A — liczba typów j. nieterminalnych w najkrótszej gramatyce
N — długość tekstu generowanego przez gramatykę (w znakach)

Jeżeli tekst jest realizacją stacjonarnego procesu stochastycznego o intensywności entropii **h** i entropii nadwyżkowej **E(N)**, to

$$\mathbf{A} \geq \mathbf{const} + [\mathbf{hN}/\log \mathbf{N}]^{1/2},$$
$$\mathbf{A} \geq \mathbf{const} + \mathbf{E(N)}/\log \mathbf{N}. \quad \textbf{(??)}$$

Intensywność entropii jest pewną miarą losowości procesu, zaś entropia nadwyżkowa jest pewną miarą jego pamięci.

Wolfgang Hilberg (1990), Frequenz, 44:243–248, w oparciu o dane Shannona (1950) przypuścił, że dla j. naturalnego $\mathbf{E(N)} \propto \mathbf{N}^{1/2}$.

Podsumowanie

- Przedstawiłem 4 modele objaśniające prawo Zipfa:
  - model „małpy przy klawiaturze” (Mandelbrot (?), Miller),
  - losowanie ze sprzężeniem zwrotnym (Simon),
  - minimalizacja średniej długości na bit informacji (Mandelbrot),
  - model najmniejszej gramatyki (Dębowski, praca w toku).
- Objaśnienia te korzystają z założeń różnej natury:
  - istnieje pewien probabilistyczny model tekstu,
  - tekst jest efektem optymalizacji pewnych kryteriów,
  - słowa w tekście może wyodrębnić pewien algorytm.

Przedstawione objaśnienia to “inspirujące potwory”.
Są zachętą do badań na pograniczu lingwistyki i matematyki.

`www.ipipan.waw.pl/~ldebowsk`

Objaśnienia mają być wstępem do teorii ilościowych

Descartes, with his vortices, his hooked atoms, and the like, explained everything and calculated nothing. Newton, with the inverse square law of gravitation, calculated everything and explained nothing.

René Thom, *Structural Stability and Morphogenesis*