Statystyczne prawa języka

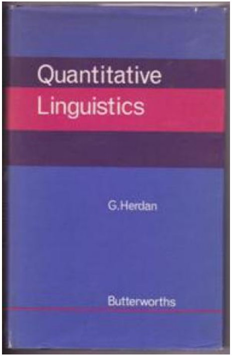
- wyjątkowo interesujące z punktu widzenia fizyków (również: matematyków, statystyków etc),

Statystyczne prawa języka

- wyjątkowo interesujące z punktu widzenia fizyków (również: matematyków, statystyków etc),
- sam Gustav Herdan (1897–1968, wybitny austriacki lingwista) stwierdził w 1964 r: *'language in use' cannot be studied without statistics*,

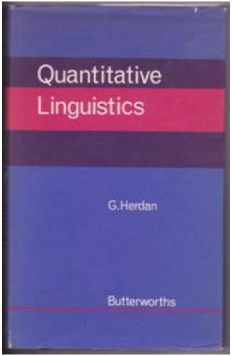
Statystyczne prawa języka

- wyjątkowo interesujące z punktu widzenia fizyków (również: matematyków, statystyków etc),
- sam Gustav Herdan (1897–1968, wybitny austriacki lingwista) stwierdził w 1964 r: *'language in use' cannot be studied without statistics*,
- ogólnie podobnymi problemami zajmuje się **lingwistyka kwantytatywna** (*quantitative linguistics*), dość interdyscyplinarny dział nauki



Statystyczne prawa języka

- wyjątkowo interesujące z punktu widzenia fizyków (również: matematyków, statystyków etc),
- sam Gustav Herdan (1897–1968, wybitny austriacki lingwista) stwierdził w 1964 r: *'language in use' cannot be studied without statistics*,
- ogólnie podobnymi problemami zajmuje się **lingwistyka kwantytatywna** (*quantitative linguistics*), dość interdyscyplinarny dział nauki



Ludek Hrebicek (2005)

...the notion `law` *(in the narrower sense scientific* `law`*) in linguistics and especially in quantitative linguistics ... need not obtain some special comprehension different from its validity in other sciences. Probably, the best delimitation of this concept can be found in the works by the philosopher of scientific knowledge Karl Raimund Popper...*

Główne prawa lingwistyki

prawo	obserwable	postać

Główne prawa lingwistyki

prawo	**obserwable**	**postać**
Zipfa	f: częstość słowa w, r: ranga słowa w	$f(r) = Ar^{-\alpha}$

Główne prawa lingwistyki

prawo	obserwable	postać
Zipfa	f: częstość słowa w, r: ranga słowa w	$f(r) = Ar^{-\alpha}$
Menzeratha-Altmanna	x: długość całości, y: rozmiar części	$y = Bx^{\beta}\mathrm{e}^{-\gamma x}$

Główne prawa lingwistyki

prawo	obserwable	postać
Zipfa	f: częstość słowa w, r: ranga słowa w	$f(r) = Ar^{-\alpha}$
Menzeratha-Altmanna	x: długość całości, y: rozmiar części	$y = Bx^{\beta}\mathrm{e}^{-\gamma x}$
Heapa	V: liczba słów, N rozmiar bazy danych	$V \sim N^{\alpha}$

Główne prawa lingwistyki

prawo	**obserwable**	**postać**
Zipfa	f: częstość słowa w, r: ranga słowa w	$f(r) = Ar^{-\alpha}$
Menzeratha-Altmanna	x: długość całości, y: rozmiar części	$y = Bx^{\beta}\mathrm{e}^{-\gamma x}$
Heapa	V: liczba słów, N rozmiar bazy danych	$V \sim N^{\alpha}$

Inne prawa

prawo	**obserwable**	**postać**
rekurencji	τ: odległość między słowami	$P(\tau) = \exp(\alpha\tau)^{\beta}$
korelacji długo-zasięgowych	$C(\tau)$: autokorelacja z przesunięciem τ	$C(\tau) \sim \tau^{-\alpha}$
skalowania entropii	H entropia tekstu z blokami o rozmairze n	$H \sim \alpha n^{\beta} + \gamma n$
Taylora	σ: odchylenie stand. wokół średniej μ	$\sigma \sim \mu^{\alpha}$

Prawo Zipfa

- **George Zipf** – amerykański lingwista (1902–1950) spopularyzował to eponimiczne prawo w latach 30-tych i 40-tych XX w., ale nie uznawał się z jego odkrywcę,

Prawo Zipfa

- **George Zipf** – amerykański lingwista (1902–1950) spopularyzował to eponimiczne prawo w latach 30-tych i 40-tych XX w., ale nie uznawał się z jego odkrywcę,
- jest jednym z wielu praw związanych z **potegowym** rozkładem prawdopobieństwa pewnych zmiennych,

Prawo Zipfa

- **George Zipf** – amerykański lingwista (1902–1950) spopularyzował to eponimiczne prawo w latach 30-tych i 40-tych XX w., ale nie uznawał się z jego odkrywcę,
- jest jednym z wielu praw związanych z **potegowym** rozkładem prawdopobieństwa pewnych zmiennych,
- historyczne sformułowanie brzmi następująco:
  Jeśli słowa (typy) otrzymają rangę (ranking) zgodnie z częstotliwością wystepowania $r = 1, 2, ..., V$, to częstość $f(r)$ r-tego słowa skaluje się z rangą jak
  $$f(r) = \frac{f(1)}{r},$$
  gdzie $f(1)$ to częstość słowa najczęstszego.
- w powyższym równaniu występuje ten problem, że stała nie jest najlepiej dobrana w przypadku dużych r, gdyż dla $f(1) > 0$ istnieje zawsze takie r^*, że $\sum_{r=1}^{r^*} f(1)/r > 1$

Prawo Zipfa

- w związku z tym, obecnie używa się po prostu postaci:

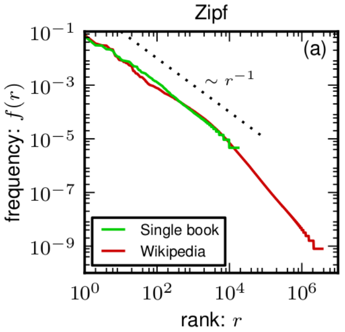
$$f(r) = \frac{\beta)}{r^{\alpha}},$$

gdzie β jest stałą normującą, a $\alpha \geq 1$,

- związek z innymi prawami potęgowymi motywuje również inną postać

$$P(f) = \frac{\beta^{+})}{f^{\alpha^{+}}},$$

gdzie $\alpha^{+} = 1 + \frac{1}{\alpha}$.



Pojedyncza książka, tu Moby Dick Melville'a, do pobrania ze strony Projektu Gutenberg `http://www.gutenberg.org/`. Czerwony – angielska Wikipedia, pobrana z `http://dumps.wikimedia.org`

Prawo Menzeratha-Altmanna

- **Paul Menzerath** (1883–1954) spopularyzował to prawo w sposób jakosciowy, odnoszący się do **fonemów**:

Prawo Menzeratha-Altmanna

- **Paul Menzerath** (1883–1954) spopularyzował to prawo w sposób jakosciowy, odnoszący się do **fonemów**:
  im większa całości, tym mniejsze jej części,

Prawo Menzeratha-Altmanna

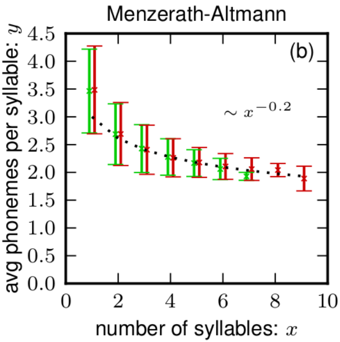
- **Paul Menzerath** (1883–1954) spopularyzował to prawo w sposób jakosciowy, odnoszący się do **fonemów**:
  im większa całości, tym mniejsze jej części,
- prace Menzeratha zostały spopularyzowane przez **Gabriela Altmanna** i przedstawione w sposób ilościowy za pomocą równania
  $$y = Bx^{\beta}\mathrm{e}^{-\gamma x},$$
  gdzie x oznacza długość całości a y rozmiar (średni) części.

Prawo Menzeratha-Altmanna

- **Paul Menzerath** (1883–1954) spopularyzował to prawo w sposób jakosciowy, odnoszący się do **fonemów**: im większa całości, tym mniejsze jej części,
- prace Menzeratha zostały spopularyzowane przez **Gabriela Altmanna** i przedstawione w sposób ilościowy za pomocą równania

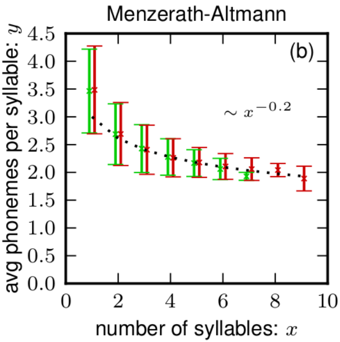
$$y = Bx^{\beta} \mathrm{e}^{-\gamma x},$$

gdzie x oznacza długość całości a y rozmiar (średni) części.



Prawo Menzeratha-Altmanna

- **Paul Menzerath** (1883–1954) spopularyzował to prawo w sposób jakosciowy, odnoszący się do **fonemów**:
  im większa całości, tym mniejsze jej części,
- prace Menzeratha zostały spopularyzowane przez **Gabriela Altmanna** i przedstawione w sposób ilościowy za pomocą równania
  $$y = Bx^{\beta} \mathrm{e}^{-\gamma x},$$
  gdzie x oznacza długość całości a y rozmiar (średni) części.



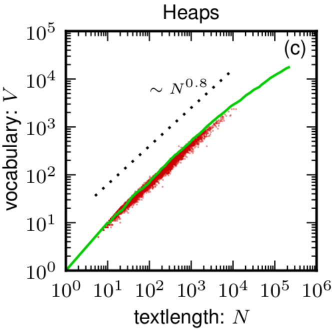
Rozmiar słowa jest mierzony przez liczbę sylab x_w, natomiast rozmiar składowych jako liczbę fonemów na sylabę $y_w = z_w / x_w$. Wartość y_w jest usredniona po wszystkich słowach w z $x_w = x$.

Prawo Heapa

- **prawo Heapa** (*Heap's law*) mówi, że:
  liczba różnych słów skaluje się potegowo wraz z całkowitą liczbą słów,
  $$V \sim N^{\alpha},$$
  gdzie V to liczba różnych słów (lub typów) a N to całkowita liczba słów.

Prawo Heapa

- **prawo Heapa** (*Heap's law*) mówi, że:
  liczba różnych słów skaluje się potegowo wraz z całkowitą liczbą słów,
  $$V \sim N^{\alpha},$$
  gdzie V to liczba różnych słów (lub typów) a N to całkowita liczba słów.

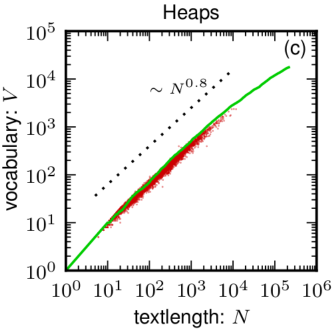


Prawo Heapa

- **prawo Heapa** (*Heap's law*) mówi, że: liczba różnych słów skaluje się potegowo wraz z całkowitą liczbą słów,

$$V \sim N^{\alpha},$$

gdzie V to liczba różnych słów (lub typów) a N to całkowita liczba słów.



- zwykle $0 < \alpha < 1$,
- dla Moby Dicka N wzrasta z każdym słowem, więc $V(N)$ tworzy krzywą,
- dla Wikipedii V i N są wyznaczone dla każdego dokumentu

Czy pomiędzy prawem Heapa oraz Zipfa można wykazać związek? Otóż tak, przy założeniu pewnego modelu zerowego *null model*:

Czy pomiędzy prawem Heapa oraz Zipfa można wykazać związek? Otóż tak, przy założeniu pewnego modelu zerowego *null model*:

- zakładamy, że słowa tworzone są zgodnie z procesem Poissona,

Czy pomiędzy prawem Heapa oraz Zipfa można wykazać związek? Otóż tak, przy założeniu pewnego modelu zerowego *null model*:

- zakładamy, że słowa tworzone są zgodnie z procesem Poissona,
- wtedy liczba różnych słów jest dana jako

  $$N(M) = \sum_r \mathrm{I}\left[n_r(M, f_r)\right],$$

  gdzie n_r to całkowita liczba przypadków wystąpienia r-tego słowa w procesie Poissona o długości M z częstością f_r, a $\mathrm{I}(..)$ to funkcja wskaźnikowa,

Czy pomiędzy prawem Heapa oraz Zipfa można wykazać związek? Otóż tak, przy założeniu pewnego modelu zerowego *null model*:

- zakładamy, że słowa tworzone są zgodnie z procesem Poissona,
- wtedy liczba różnych słów jest dana jako

  $$N(M) = \sum_r \mathrm{I}\left[n_r(M, f_r)\right],$$

  gdzie n_r to całkowita liczba przypadków wystąpienia r-tego słowa w procesie Poissona o długości M z częstością f_r, a $\mathrm{I}(..)$ to funkcja wskaźnikowa,
- uśrednianie po realizacjach procesu Poissona daje

  $$\mathrm{E}[\mathrm{I}\left[n_r(M, f_r)\right]] = 1 - \exp(-Mf_r),$$

  co pokazuje prawdopodobieństwo, że słowo o randze r pojawia się co najmniej raz w tekście o długości M,
- dla wszystkich słów

  $$\mathrm{E}[N(M)] \equiv \mu(M) = \sum_r 1 - \exp(-Mf_r),$$

- w podobny sposób można wyznaczyć wariancję,

- w podobny sposób można wyznaczyć wariancję,

$$\mathrm{V}[N(M)] \equiv \sigma(M)^2 \equiv \mathrm{V}[N(M)^2] - \mathrm{V}[N(M)]^2 = \\ = \sum_r \exp(-Mf_r) - \exp(-2Mf_r)$$

- jeśli w tym momencie założymy postać prawa Zipfa, to dla $M \gg 1$ (po długich obliczniach) otrzymamy

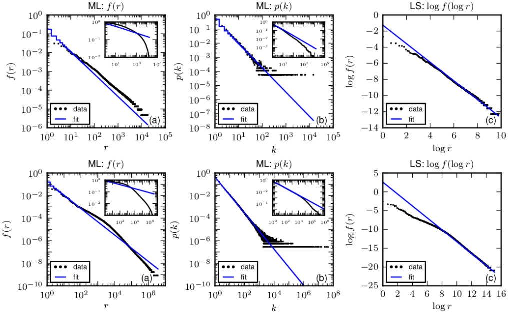
$$\mathrm{E}[N(M)] \sim M^\lambda,$$

gdzie $\alpha = 1/\lambda$, natomiast z relacji związanej z wariancją dostaniemy

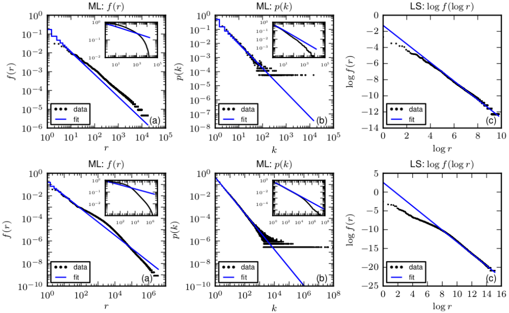
$$\sigma(M)^2 \sim \mu^\beta,$$

z $\beta = 1/2$.

Dopasowywanie wykładnika



Dopasowywanie wykładnika



Book	Rank: $f(r)$		Frequency: $P(f)$		Linear: $\log f(\log r)$	
	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	R^2
Alice's Adventures in Wonderland (L. Carroll)	1.22	$< 10^{-4}$	1.46	$< 10^{-4}$	1.21	0.97
The Voyage Of The Beagle (C. Darwin)	1.20	$< 10^{-4}$	1.59	$< 10^{-4}$	1.29	0.97
The Jungle (U. Sinclair)	1.21	$< 10^{-4}$	1.45	$< 10^{-4}$	1.22	0.98
Life On The Mississippi (M. Twain)	1.20	$< 10^{-4}$	1.38	$< 10^{-4}$	1.16	0.98
Moby Dick; or The Whale (H. Melville)	1.19	$< 10^{-4}$	1.38	$< 10^{-4}$	1.15	0.98
Pride and Prejudice (J. Austen)	1.21	$< 10^{-4}$	1.66	$< 10^{-4}$	1.35	0.98
Don Quixote (M. Cervantes)	1.21	$< 10^{-4}$	1.70	$< 10^{-4}$	1.38	0.98
The Adventures of Tom Sawyer (M. Twain)	1.21	$< 10^{-4}$	1.29	$< 10^{-4}$	1.12	0.98
Ulysses (J. Joyce)	1.18	$< 10^{-4}$	1.15	$< 10^{-4}$	1.03	0.97
War and Peace (L. Tolstoy)	1.20	$< 10^{-4}$	1.84	$< 10^{-4}$	1.44	0.97
English Wikipedia	1.17	$< 10^{-4}$	1.60	$< 10^{-4}$	1.58	0.99

Brak uwzględnienia korelacji