Poprawki do praw Zipfa i Heapsa

oparte na modelowaniu frakcji hapaksów

Łukasz Dębowski

Instytut Podstaw Informatyki PAN

Seminarium IPI PAN, 13 listopada 2023

Prawa Zipfa i Heapsa

Prawo Zipfa:

- najsłynniejsze prawo językoznawstwa ilościowego (lingwistyki kwantytatywnej);
- głosi, że ***n***-te co do częstości słowo w tekście pojawia się około ***n*** razy rzadziej niż słowo najczęstsze (ranga słowa jest odwrotnie proporcjonalna do jego częstości);
- zaobserwowali je Estoup (1916), Condon (1928) i Zipf (1935);
- podobne rozkłady potęgowe obserwuje się w ekologii, socjologii, ekonomii i fizyce (Zipf, 1949).

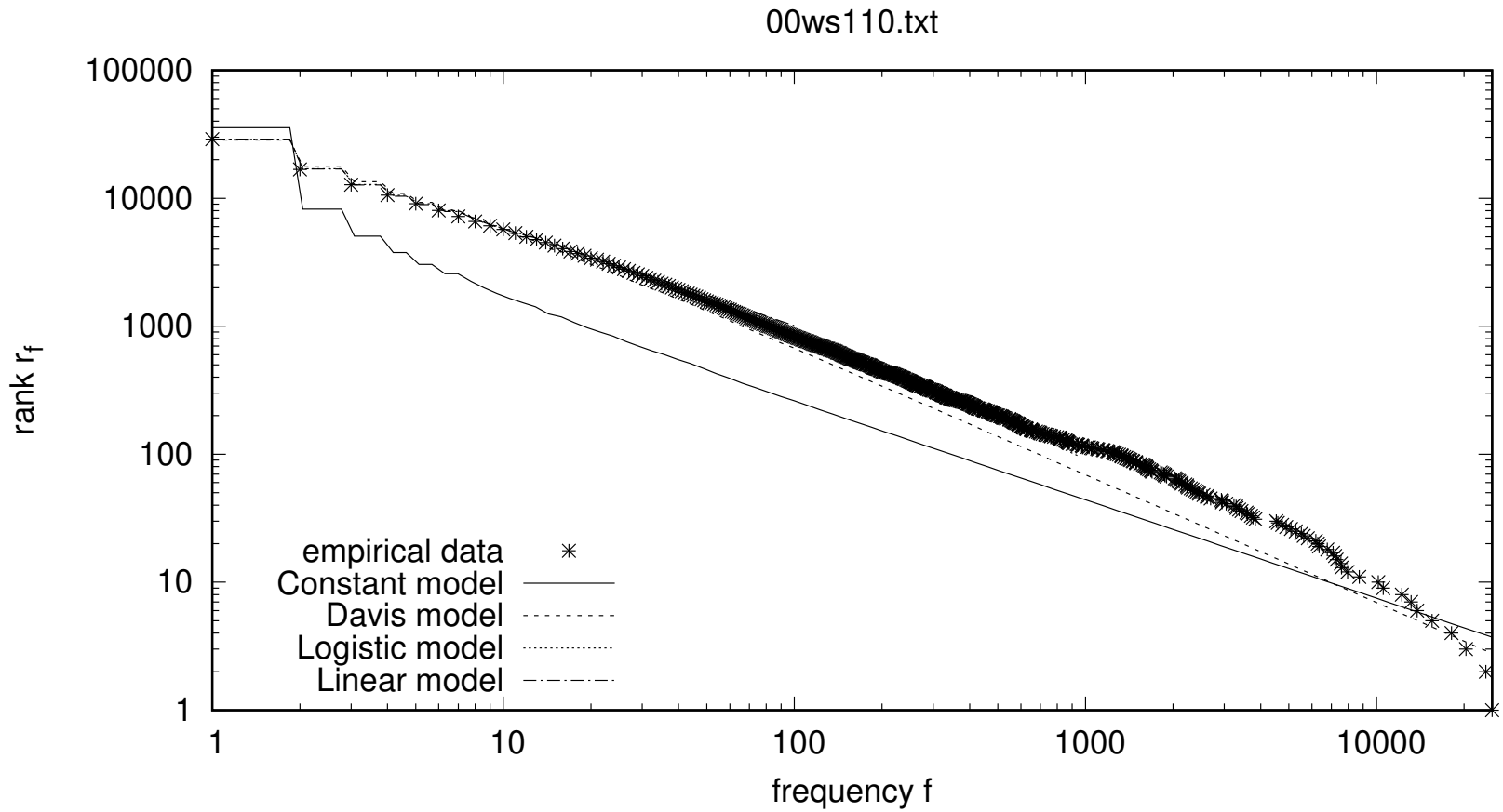
Prawo Heapsa:

- powinno być wnioskiem z prawa Zipfa (ale nie jest!);
- głosi, że liczba różnych słów w tekście rośnie w przybliżeniu jak potęga długości tekstu;
- zaobserwowali je Kuraszkiewicz i Łukaszewicz (1951), Herdan (1964) i Heaps (1978).

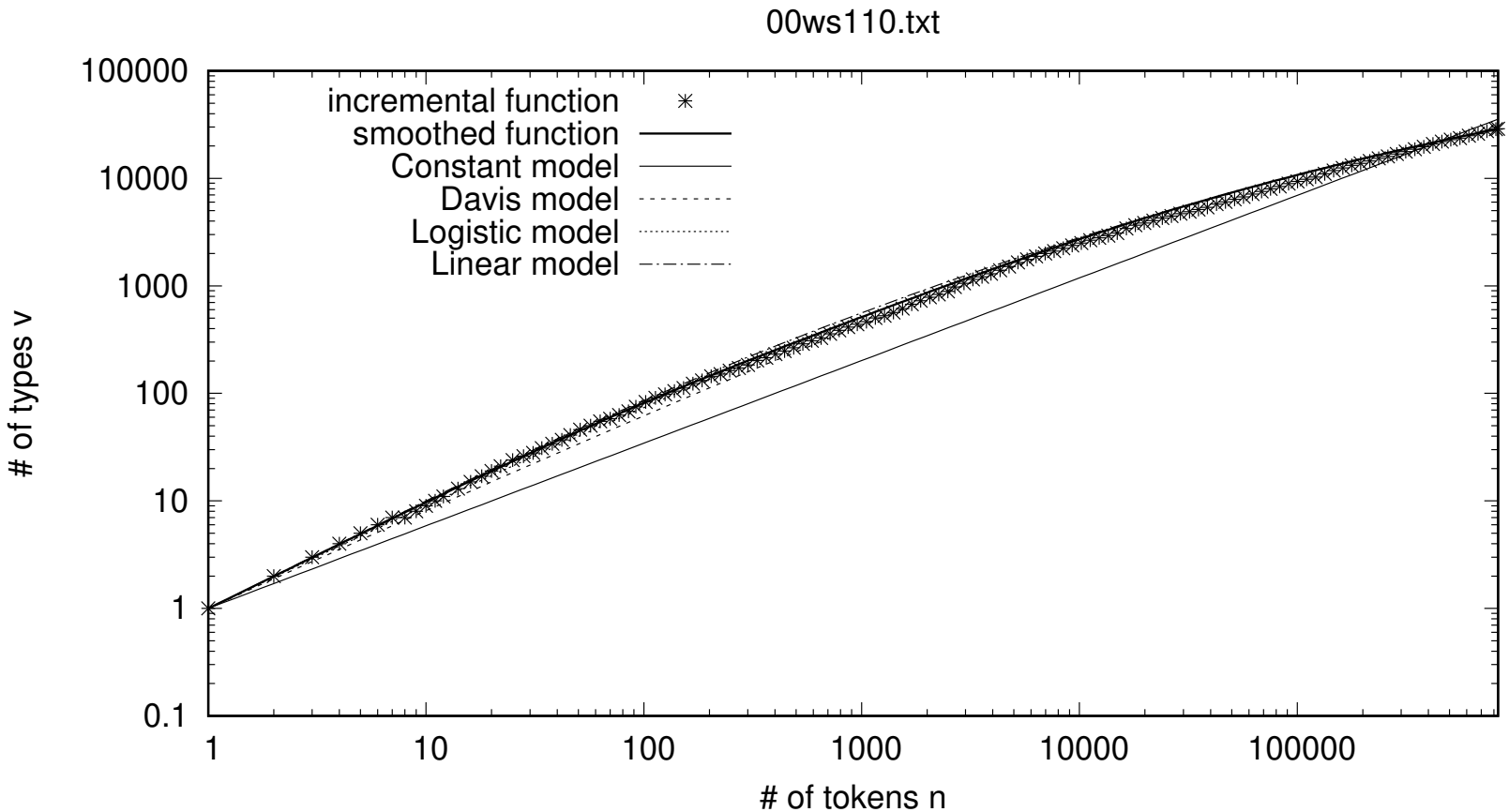
Lista frekwencyjna dla sztuk Williama Szekspira

ranga r	częstość F_r	słowo
1	21557	I
2	19059	and
3	16571	to
4	14921	of
5	14491	a
6	12077	my
7	10463	you
8	9789	in
9	8754	is
10	7428	that
...	...	...

Wykres ranga-częstość z zamienionymi osiami



Liczba różnych słów w tekście



Skąd te prawa?

Prawo Zipfa próbowano wyjaśniać na wiele sposobów:

- model małpy przy klawiaturze (Mandelbrot, 1954; Miller, 1957),
- preferencyjne przyłączanie, czyli zasada św. Mateusza (Simon, 1955),
- teoria gier i kodowanie (Harremoës i Topsøe, 2005),
- semantyka, nieergodyczność i kodowanie uniwersalne (Dębowski, 2021),
- deterministyczne sekwencje multiperiodyczne (Dębowski, 2023).

Każde z tych objaśnień otwiera nowe pola dalszej eksploracji.

Losowość w opisie zjawisk językowych jest obecna.
Pytanie, ile jej dokładnie jest i jakie jest jej rozmieszczenie.

Ograniczenie harmoniczne

Niech $F_1 \geq F_2 \geq F_3 \geq \ldots$ oraz $\sum_{r=1}^{\infty} F_r = n$.

Wówczas $r \cdot F_r \leq F_1 + F_2 + \ldots + F_r \leq n$, czyli

$$F_r \leq \frac{n}{r}. \tag{1}$$

Prawo Zipfa głosi, że

$$F_r \approx \frac{V}{r}, \tag{2}$$

gdzie V to liczba różnych słów w tekście.

Prawo Zipfa to rozkład ranga-częstość o najgrubszym ogonie.

Prawo Heapsa jest sprzeczne z prawem Zipfa!

Prawo Heapsa głosi, że

$$V = n^{\beta}, \quad 0 < \beta < 1. \tag{3}$$

Z prawa Zipfa wynika natomiast

$$n = \sum_{r=1}^{V} F_r \approx \sum_{r=1}^{V} \frac{V}{r} \approx V \log V. \tag{4}$$

Coś tu jest nie tak!!! Popełniamy pewien błąd modelowania!!!

Przedmiot tej prezentacji

Chciałbym przedstawić:

- ... stosunkowo prosty a zarazem stosunkowo dokładny ...
- ... parametryczny model ...
- ... brzegowego rozkładu częstości słów ...
- ... w tekstach dowolnej długości.

Dwa założenia modelu:

1. model urnowy
   (Khmaladze, 1988; Baayen, 2001; Milička, 2009; Davis, 2018):
   Rozkład brzegowy słów w tekście wygląda, jakby słowa były losowane na ślepo bez zwracania z pewnej urny ze słowami.
2. analityczna frakcja hapaksów (nowe!):
   Frakcja słów, które pojawiają się jeden raz, daje się przybliżyć jako prosta analityczna funkcja długości tekstu.

`https://arxiv.org/abs/2307.12896`

Model urnowy

Widmo częstości

Prawa Zipfa i Heapsa można wyznaczyć z widma częstości:

- $V(n)$ — liczba różnych słów w tekście długości n;
- $V_k(n)$ — liczba różnych słów o k wystąpieniach w tekście długości n.

Niech $R_f(n)$ — liczba różnych słów o $\geq f$ wystąpieniach:

$$R_f(n) = V(n) - \sum_{k=1}^{f-1} V_k(n). \tag{5}$$

Wykres $R_f(n)$ to wykres ranga-częstość z zamienionymi osiami.

Niech F_w — liczba wystąpień słowa w w tekście długości n:

$$V(n) = \sum_{w \in \mathbb{W}} 1\{F_w > 0\} = \sum_{w \in \mathbb{W}} [1 - 1\{F_w = 0\}], \tag{6}$$

$$V_k(n) = \sum_{w \in \mathbb{W}} 1\{F_w = k\}. \tag{7}$$

Rozkład Bernoulliego

Dla uproszczenia wzorów załóżmy, że mamy ustalony rozkład prawdopodobieństwa na słowach $w \in \mathbb{W}$,

$$p_w \geq 0, \qquad \sum_{w \in \mathbb{W}} p_w = 1. \tag{8}$$

Liczba wystąpień F_w słowa w w tekście długości n spełnia

$$P(F_w = k) = \binom{n}{k} p_w^k (1 - p_w)^{n-k}. \tag{9}$$

Oczekiwane widmo częstości wynosi

$$\mathbb{E}\, V(n) = \sum_{w \in \mathbb{W}} [1 - P(F_w = 0)] = \sum_{w \in \mathbb{W}} [1 - (1 - p_w)^n], \tag{10}$$

$$\mathbb{E}\, V_k(n) = \sum_{w \in \mathbb{W}} P(F_w = k) = \sum_{w \in \mathbb{W}} \binom{n}{k} p_w^k (1 - p_w)^{n-k}. \tag{11}$$

Przybliżenie rozkładem Poissona

Załóżmy, że p_w jest małe, a n jest duże. Wówczas

$$P(F_w = k) = \binom{n}{k} p_w^k (1-p_w)^{n-k} \approx \frac{[np_w]^k}{k!} e^{-np_w}. \quad (12)$$

Oczekiwane widmo częstości można przybliżyć jako

$$\mathbb{E}\, V(n) = \sum_{w \in \mathbb{W}} [1 - (1-p_w)^n] \approx \sum_{w \in \mathbb{W}} [1 - e^{-np_w}], \quad (13)$$

$$\mathbb{E}\, V_k(n) = \sum_{w \in \mathbb{W}} \binom{n}{k} p_w^k (1-p_w)^{n-k} \approx \sum_{w \in \mathbb{W}} \frac{[np_w]^k}{k!} e^{-np_w}. \quad (14)$$

Odtwarzanie widma częstości

Czy znając wyłącznie, jak liczba różnych słów rośnie z długością tekstu, jesteśmy w stanie odtworzyć widmo częstości?

Tak! Rożniczkując liczbę różnych słów otrzymujemy widmo:

$$\begin{aligned}\frac{(-n)^k}{k!}\frac{d^k}{dn^k}\,\mathbb{E}\,V(n) &\approx \frac{(-n)^k}{k!}\frac{d^k}{dn^k}\sum_{w\in\mathbb{W}}[1-e^{-np_w}]\\ &= \sum_{w\in\mathbb{W}}\frac{[np_w]^k}{k!}e^{-np_w}\\ &\approx \mathbb{E}\,V_k(n). \qquad (15)\end{aligned}$$

Frakcja hapaksów

(coś nowego)

Frakcja hapaksów

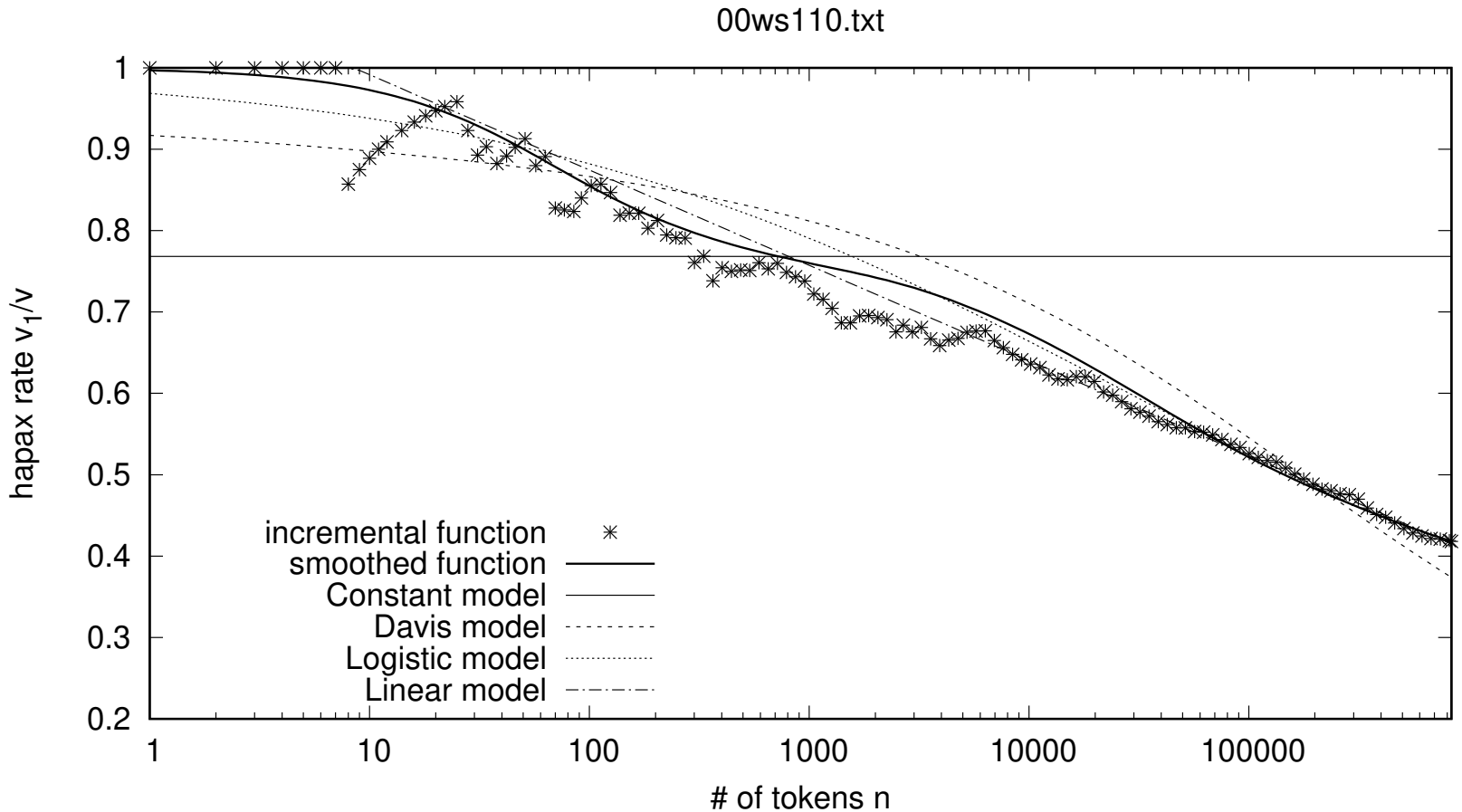
Hapaksy to słowa, które pojawiają się tylko jeden raz
(z greckiego „hapax legomena" = „raz przeczytane").

Frakcja hapaksów to ich udział w całym słowniku tekstu:

$$\boldsymbol{h(u)} := \frac{\mathbb{E}\, \boldsymbol{V_1(\exp u)}}{\mathbb{E}\, \boldsymbol{V(\exp u)}}. \tag{16}$$

Zmienna $\boldsymbol{u = \log n}$ to dogodny argument funkcji $\boldsymbol{h(u)}$,
gdyż frakcja hapaksów słabo zależy od długości tekstu.

Frakcja hapaksów dla sztuk Williama Szekspira



Frakcja hapaksów i liczba różnych słów

Z definicji mamy

$$h(\log n) = n\frac{d\,\mathbb{E}\,V(n)}{dn} : \mathbb{E}\,V(n). \tag{17}$$

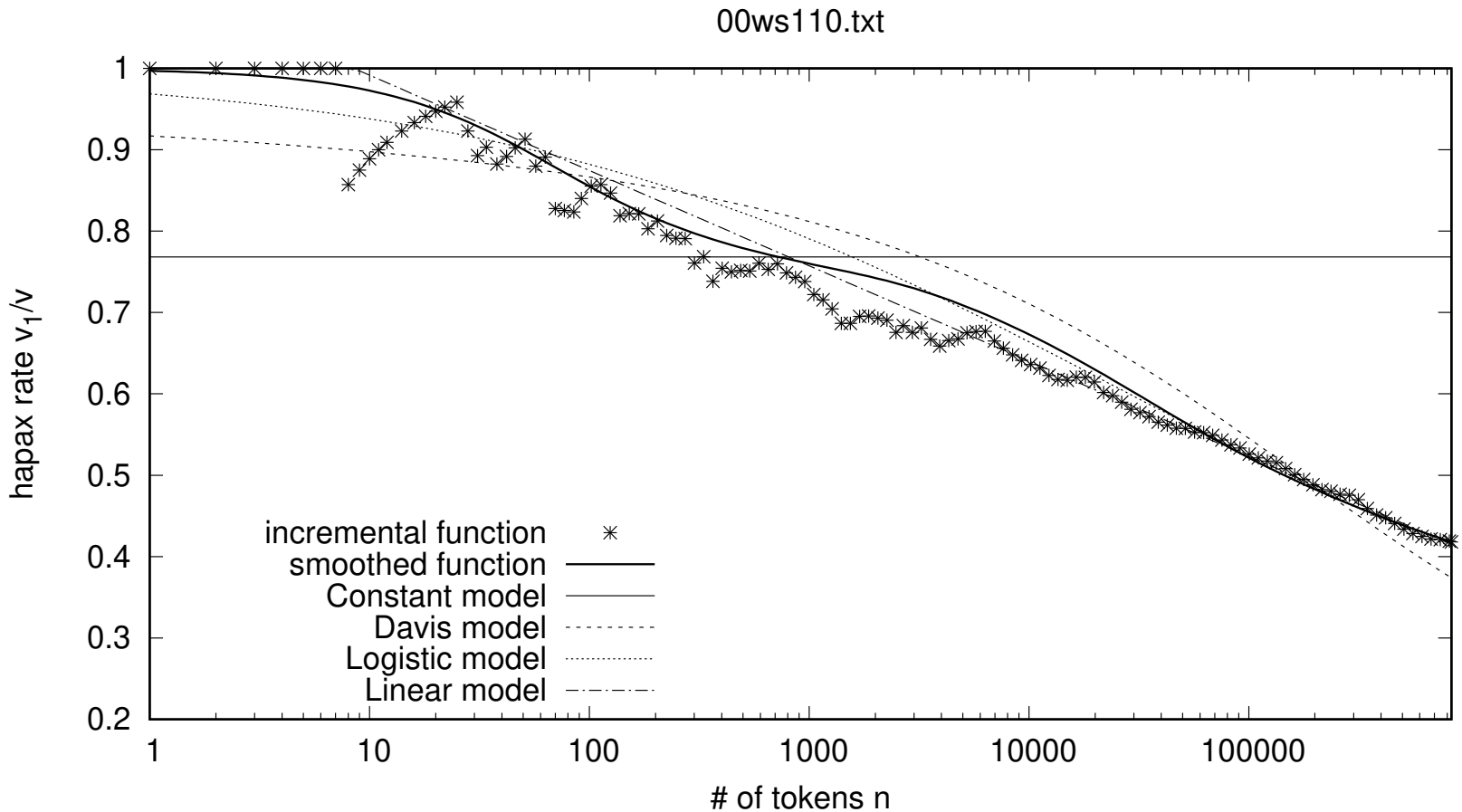
Rozwiązując to równanie różniczkowe, otrzymujemy

$$\mathbb{E}\,V(n) = \exp\left(\int_0^{\log n} h(u)du\right). \tag{18}$$

Czyli całkując funkcję $h(u)$, możemy policzyć widmo częstości.

Potencjalny słownik jest skończony, jeżeli pole pod wykresem frakcji hapaksów w skali logarytmicznej jest skończone!

Jeszcze raz rzućmy okiem!



Modele frakcji hapaksów

Cztery modele frakcji hapaksów

1. Model stały:

$$h_\beta(u) = \beta. \tag{19}$$

2. Model Davisa:

$$h_\alpha(u) = \frac{1}{u-\alpha} - \frac{1}{e^{u-\alpha}-1}. \tag{20}$$

3. Model liniowy:

$$h_{\alpha\gamma}(u) = \begin{cases} 1, & u < \alpha, \\ 1-\gamma(u-\alpha), & \alpha \le u \le \gamma^{-1}+\alpha, \\ 0, & u > \gamma^{-1}+\alpha. \end{cases} \tag{21}$$

4. Model logistyczny:

$$h_{\alpha\beta\gamma}(u) = \frac{1-\beta}{1+e^{\gamma(u-\alpha)}} + \beta. \tag{22}$$

Model 1: Model stały

Model stały zakłada, że frakcja hapaksów jest stała,

$$h(u) = \beta \in (0, 1). \tag{23}$$

Wówczas liczba różnych słów spełnia prawo Herdana-Heapsa

$$\mathbb{E}\, V(n) \approx \exp\left(\int_0^{\log n} \beta du\right) = n^{\beta}. \tag{24}$$

Mamy $\mathbb{E}\, V_k(n) \approx n^{\beta}\left(\frac{\beta}{k}\right)\prod_{i=1}^{k-1}\left(1-\frac{\beta}{i}\right)$ oraz

$$\mathbb{E}\, R_f(n) \approx n^{\beta}\prod_{i=1}^{f-1}\left(1-\frac{\beta}{i}\right). \tag{25}$$

Unormowana funkcja $\frac{\mathbb{E}\, R_f(n)}{\mathbb{E}\, V(n)}$ nie zależy od długości tekstu n.

Dla $f \to \infty$ (25) dąży do prawa Zipfa-Mandelbrota $\mathbb{E}\, R_f \propto \frac{1}{f^{\beta}}$.

Model 2: Model Davisa

Model Davisa zakłada funkcję sigmoidalną

$$h(u) = \frac{1}{u} - \frac{1}{e^u - 1}. \tag{26}$$

To implikuje logarytmiczny wzrost liczby różnych słów,

$$\mathbb{E}\, V(n) \approx \frac{n \log n}{n-1} \approx \log n, \tag{27}$$

prawo Lotki $\mathbb{E}\, V_k(1) \approx \frac{1}{k(k+1)}$ i prawo Zipfa $\mathbb{E}\, R_f(1) \approx \frac{1}{f}$.

W ogólności funkcja częstość-ranga wyraża się jako

$$\begin{aligned} \mathbb{E}\, R_f(n) &\approx \frac{\log n - \sum_{j=1}^{f-1} (1-1/n)^j / j}{(1-1/n)^f} \\ &= \sum_{j=0}^{\infty} \frac{(1-1/n)^j}{j+f} \approx \exp\left(\frac{f}{n}\right) \Gamma\left(0, \frac{f}{n}\right). \end{aligned} \tag{28}$$

Modele 3 i 4: Liniowy i logistyczny

- Model liniowy ma postać

$$h(u) = \begin{cases} 1, & u < \alpha, \\ 1 - \gamma(u - \alpha), & \alpha \leq u \leq \gamma^{-1} + \alpha, \\ 0, & u > \gamma^{-1} + \alpha. \end{cases} \tag{29}$$

  Implikuje on skończony słownik (ok. 20 000 słów).

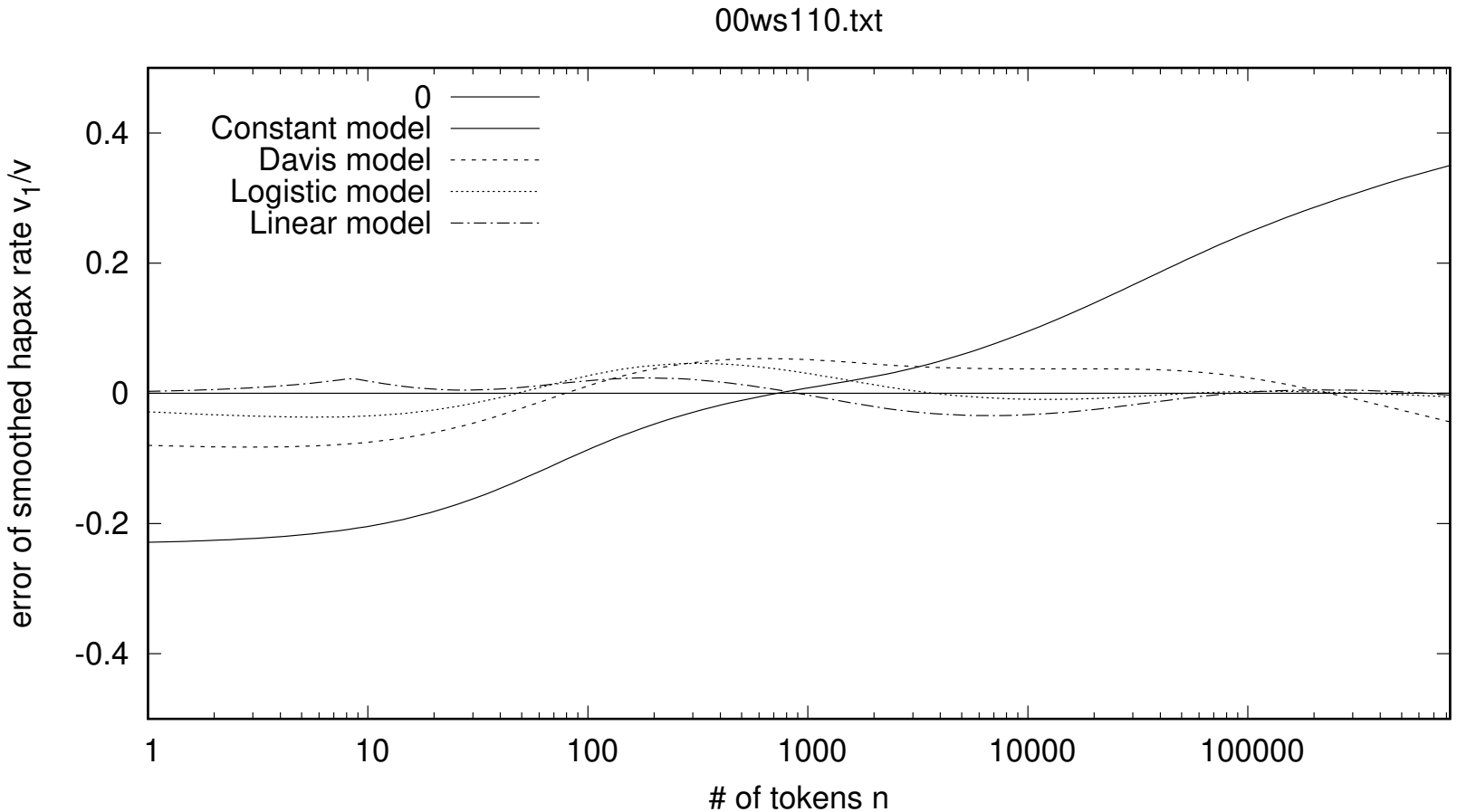
- Model logistyczny ma postać

$$h(u) = \frac{1 - \beta}{1 + e^{\gamma(u-\alpha)}} + \beta. \tag{30}$$

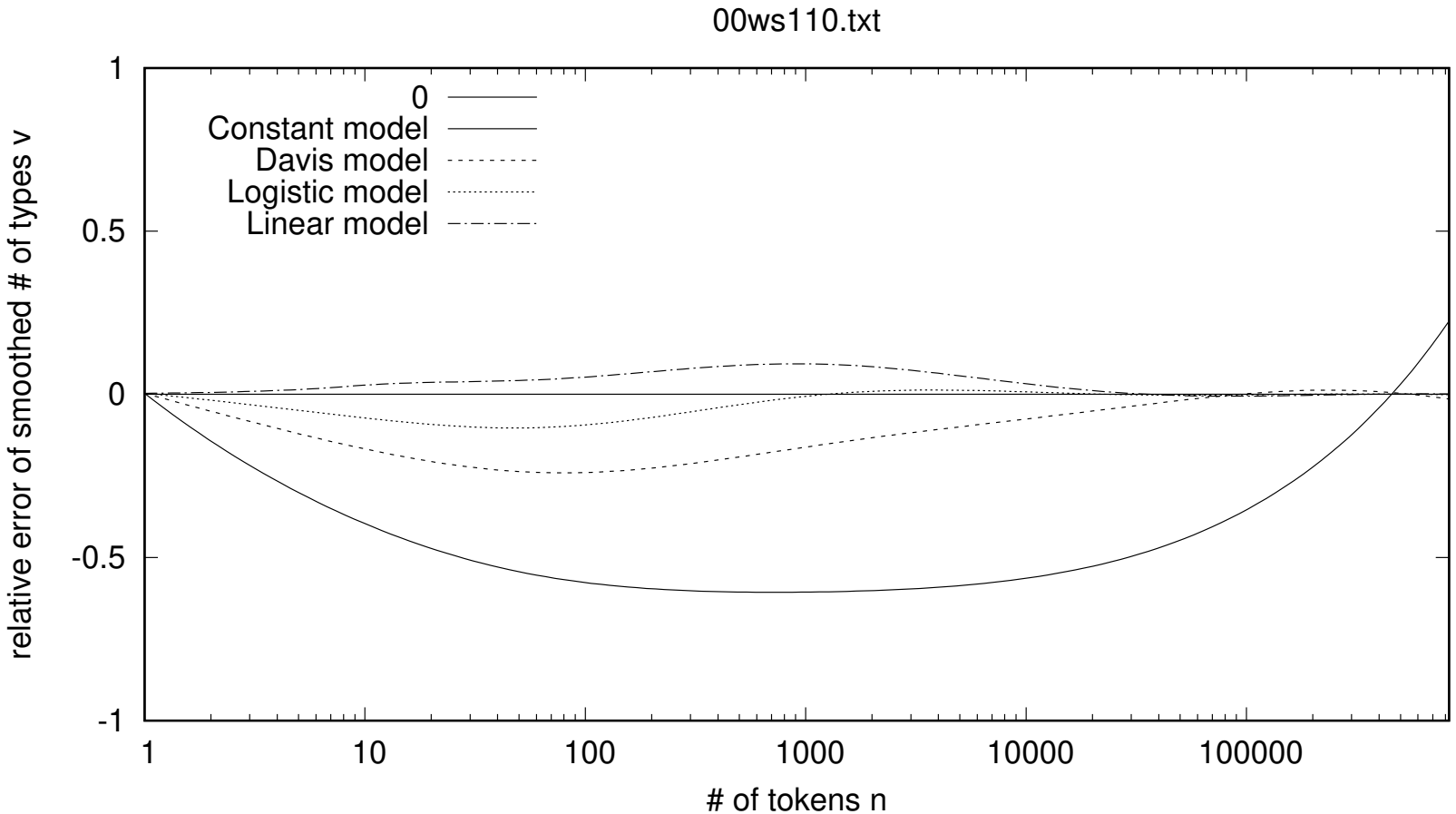
  Dla $\beta = 0$ także implikuje skończony słownik.

W obu przypadkach widmo częstości daje się policzyć za pomocą rekursyjnie zdefiniowanych wielomianów.

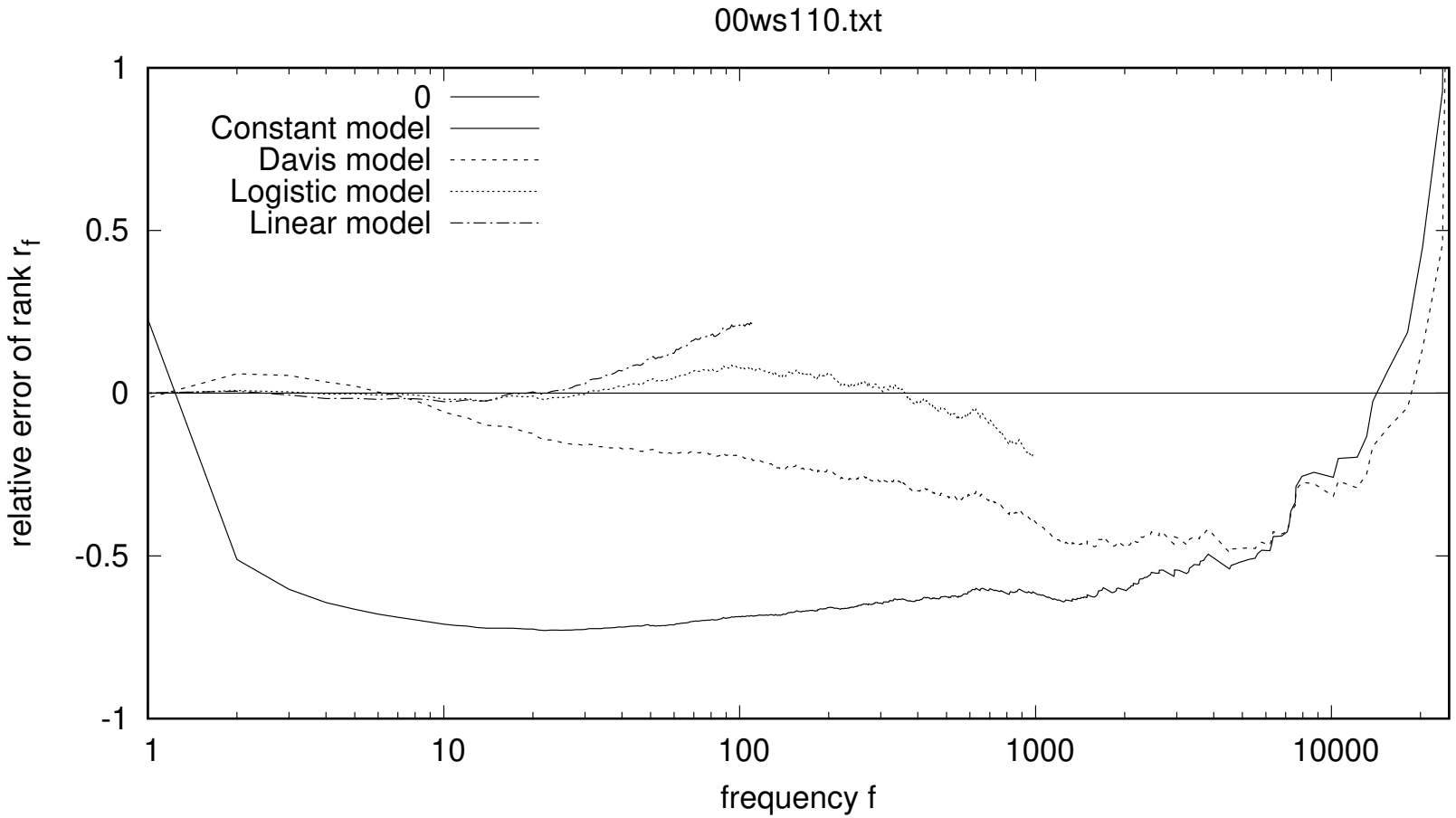
Błąd przewidywanej frakcji hapaksów



Błąd przewidywanej liczby różnych słów



Błąd przewidywanej funkcji częstość-ranga



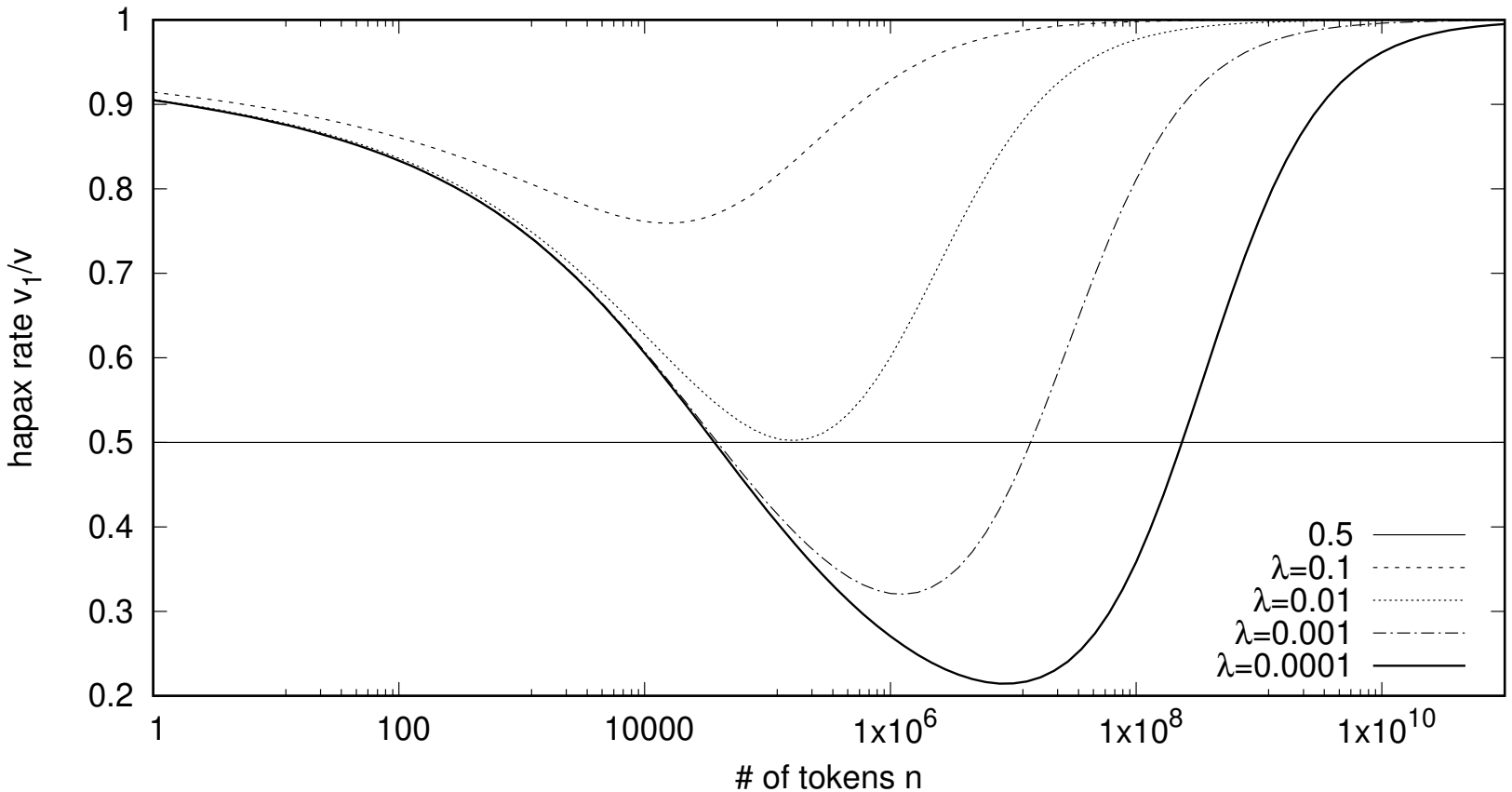
Dopasowane parametry — Projekt Gutenberg (j. angielski)

plik	stały	Davisa	logisticzny			liniowy		długość
	β	α	γ	β	α	γ	α	N
00ws110.txt	0.768	12.06	0.314	0.218	10.11	0.0509	2.14	835726
1ours10.txt	0.797	11.55	0.318	0.203	9.72	0.0507	1.7	128963
2000010.txt	0.801	11.48	0.323	0.008	10.62	0.0578	2.22	101247
2cahe10.txt	0.796	12.12	0.314	0	11.38	0.0576	2.79	298339
5wiab10.txt	0.808	11.64	0.315	0.001	10.86	0.0552	2.13	92558
800lg10.txt	0.799	11.43	0.327	0.162	9.77	0.0534	1.84	95493
csnva10.txt	0.732	11.39	0.308	0.157	9.94	0.0542	1.87	1268149
dbrry10.txt	0.787	11.39	0.325	0.065	10.31	0.0583	2.23	159710
dscmn10.txt	0.774	11.5	0.328	0	10.75	0.0629	2.71	312075
gltrv10.txt	0.796	11.4	0.322	0.001	10.62	0.0584	2.22	104909
milnd10.txt	0.773	11.14	0.347	0.127	9.63	0.0608	2.24	195064
mt7bg10.txt	0.775	11.91	0.296	0.001	11.45	0.0565	2.55	519886
stlla10.txt	0.757	10.91	0.333	0.231	8.87	0.0536	1.45	245882
wmcry10.txt	0.799	11.69	0.314	0	10.96	0.0567	2.34	145487
średnia	0.783	11.54	0.32	0.084	10.36	0.0562	2.17	321678

Drugi reżim dla wielkich korpusów

Fengxiang (2010) zauważył U-kształny wykres frakcji hapaksów.

Możemy to zamodelować jako kombinację wypukłą modelu Davisa z $\alpha = 10.51$ i modelu stałego z $\beta = 1$:



Konkluzja

Wykres ranga-częstość z zamienionymi osiami łatwiej analizować.

- Zakładając prostą postać funkcyjną frakcji hapaksów, jesteśmy w stanie dość dokładnie zamodelować liczbę różnych słów i wykres ranga-częstość dla tekstów dowolnej długości.
- Te poprawki do praw Zipfa i Herdana mają raptem 2–3 parametry, ale są znacznie dokładniejsze.
- Dla najdokładniejszego modelu logistycznego wykres ranga-częstość daje się policzyć za pomocą rekursyjnie zdefiniowanych wielomianów.
- Wykres frakcji hapaksów sugeruje, że słownik jest skończony (ok. 20 000 słów), ale trend odwraca się dla wielkich korpusów.

Wykres frakcji hapaksów to narzędzie diagnostyczne!

Literatura I

R. H. Baayen (2001) *Word frequency distributions*, Dordrecht: Kluwer Academic Publishers.

E. U. Condon (1928) *Statistics of vocabulary*, Science, t. 67, nr 1733, s. 300–300.

V. Davis (2018) *Types, Tokens, and Hapaxes: A New Heap's Law*, Glottotheory, t. 9, nr 2, s. 113–129.

Ł. Dębowski (2021) *Information Theory Meets Power Laws: Stochastic Processes and Language Models*, New York: Wiley & Sons.

— (2023) *A Simplistic Model of Neural Scaling Laws: Multiperiodic Santa Fe Processes*. `https://arxiv.org/abs/2302.09049`.

J. B. Estoup (1916) *Gammes sténographiques*, Paris: Institut Stenographique de France.

Literatura II

F. Fengxiang (2010) *An Asymptotic Model for the English Hapax/Vocabulary Ratio*, Computational Linguistics, t. 36, nr 4, s. 631–637.

P. Harremoës, F. Topsøe (2005) *Zipf's law, hyperbolic distributions and entropy loss*, Electronic Notes in Discrete Mathematics, t. 21, s. 315–318. General Theory of Information Transfer and Combinatorics.

H. S. Heaps (1978) *Information Retrieval—Computational and Theoretical Aspects*, New York: Academic Press.

G. Herdan (1964) *Quantitative Linguistics*, London: Butterworths.

E. Khmaladze (1988) *The statistical analysis of large number of rare events*. Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam.

Literatura III

W. Kuraszkiewicz, J. Łukaszewicz (1951) *Ilość różnych wyrazów w zależności od długości tekstu*, Pamiętnik Literacki, t. 42(1), s. 168–182.

B. Mandelbrot (1954) *Structure formelle des textes et communication*, Word, t. 10, s. 1–27.

J. Milička (2009) *Type-token & Hapax-token Relation: A Combinatorial Model*, Glottotheory, t. 2, nr 1, s. 99–110.

G. A. Miller (1957) *Some effects of intermittent silence*, American Journal of Psychology, t. 70, s. 311–314.

H. A. Simon (1955) *On a class of skew distribution functions*, Biometrika, t. 42, s. 425–440.

G. K. Zipf (1935) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Boston: Houghton Mifflin.

— (1949) *Human Behavior and the Principle of Least Effort*, Reading: Addison-Wesley.