Principled Analytic Corrections of Zipf's Law
(that Stem from Simple Hapax Rate Models)

Łukasz Dębowski [1] Iván González Torre [2,3]

[1]Institute of Computer Science, Polish Academy of Sciences

[2]Language and Speech Laboratory,
Universidad del País Vasco/Euskal Herriko Unibertsitatea

[3]Departamento de Matemática Aplicada, Universidad Politécnica de Madrid

QUALICO 2023, Lausanne, June 28–30

Introduction

- We will derive corrections to Zipf's law for texts of any size.
- Our derivation rests on two assumptions:
  1. The first assumption is the urn model which states that word frequency distributions look as if the word tokens were generated by a memoryless source.
  2. The second assumption is that we have an exact analytic formula for the hapax rate function.
- Assumption 1 was developed by Khmaladze (1988) and Baayen (2001). Milička (2009) and Davis (2018) found it out later independently.

Our contribution is Assumption 2 and a few new formulae.
We cautiously hope that it is not a re-discovery.

Notation

- Fix a text $T = (T_1, T_2, ..., T_n)$.
- Let $1\{\text{true}\} := 1$ and $1\{\text{false}\} := 0$.
- The frequency of word w is $F(w) := \sum_{i=1}^{n} 1\{T_i = w\}$.
- The number of types with frequency k is $V_k := \sum_{w:F(w)=k} 1$.
- The frequency spectrum is sequence $(V_1, V_2, ...)$.
- The number of types is $V = \sum_{w:F(w)>0} 1 = \sum_{k=1}^{\infty} V_k$.
- The number of tokens is $n = \sum_{w:F(w)>0} F(w) = \sum_{k=1}^{\infty} kV_k$.
- The inverse rank-frequency function is $R_f = V - \sum_{k=1}^{f-1} V_k$.

Expected frequency spectrum

Let p_w be the probability of word w for a memoryless source.

According to the urn model (Khmaladze, 1988; Baayen, 2001; Milička, 2009; Davis, 2018), the expected number of types and the expected frequency spectrum for the text length n are

$$\mathbb{E}\, V = \sum_w [1 - (1 - p_w)^n] \approx g(n) := \sum_w [1 - e^{-np_w}],$$

$$\mathbb{E}\, V_k = \sum_w \binom{n}{k} p_w^k (1 - p_w)^{n-k} \approx g(n|k) := \sum_w \frac{[np_w]^k}{k!} e^{-np_w},$$

where $\binom{n}{k} := \dfrac{n!}{k![n-k]!} \approx \dfrac{n^k}{k!}$ for $k \ll n$.

Expected inverse rank-frequency function

As observed by Baayen (2001) and Davis (2018), the frequency spectrum $g(n|k)$ can be evaluated by taking derivatives of the vocabulary size function $g(n)$.

In particular, for a given function $g(n)$, we may evaluate the expected inverse rank-frequency function as

$$\mathbb{E}\, R_f = \mathbb{E}\, V - \sum_{k=1}^{f-1} \mathbb{E}\, V_k \approx g(n||f) := \underbrace{g(n) + \sum_{k=1}^{f-1} \frac{(-n)^k}{k!} \frac{d^k g(n)}{dn^k}}_{\text{truncated Taylor series for } g(0)}.$$

Zipf's law plot with swapped axes is easier to analyze!

The hapax rate function

The fraction of words that occur exactly once is

$$\frac{\mathbb{E}\, V_1}{\mathbb{E}\, V} \approx h(\log n) := \frac{g(n|1)}{g(n)}.$$

Variable $u = \log n$ is a natural choice of the argument for the hapax rate function $h(u)$. We have

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

Function $h(u)$ is well-defined if it is analytic and satisfies conditions

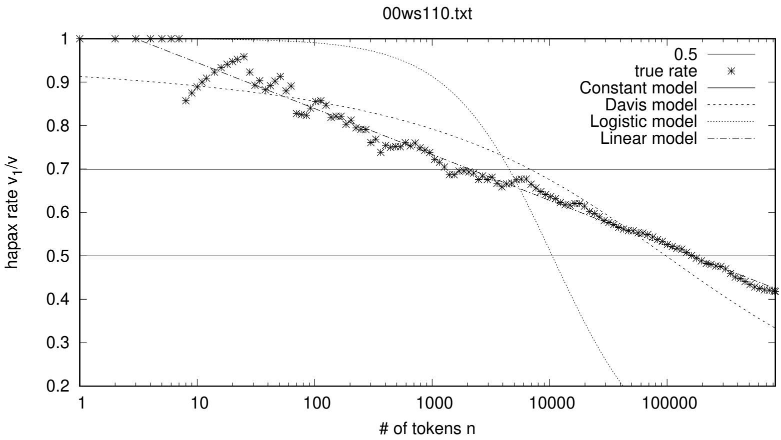
$$\int_{-\infty}^{0} h(u)du = \infty, \quad h(u|k) \geq 0 \text{ for } k \geq 1,$$

where we define recursively $h(u|0) := -1$ and

$$h(u|k) := \left[1 - \frac{1}{k}\left(1 + h(u) + \frac{d}{du}\right)\right] h(u|k-1), \quad k \geq 1.$$

Empirical hapax rate

The hapax rate for Shakespeare's First Folio.



Model 1: Constant model

The constant model assumes a constant hapax rate function,

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

Then the vocabulary size function follows Herdan-Heaps' law

$$\mathbb{E}\, V \approx g(n) = n^\beta. \tag{2}$$

We have $\mathbb{E}\, V_k \approx g(n|k) = n^\beta \left(\frac{\beta}{k}\right) \prod_{i=1}^{k-1} \left(1 - \frac{\beta}{i}\right)$ and

$$\mathbb{E}\, R_f \approx g(n||f) = n^\beta \prod_{i=1}^{f-1} \left(1 - \frac{\beta}{i}\right). \tag{3}$$

In this case, normalized ranks $\frac{\mathbb{E}\, R_f}{\mathbb{E}\, V}$ do not depend on text size n.

For $f \to \infty$, (3) tends to Zipf-Mandelbrot's law $\mathbb{E}\, R_f \propto \frac{1}{f^\beta}$.

Model 2: Davis model

The Davis model is the sigmoid hapax rate function of form

$$h(u) = \frac{1}{u} - \frac{1}{e^u - 1}. \tag{4}$$

This implies a logarithmic growth of the vocabulary,

$$\mathbb{E}\, V \approx g(n) = \frac{n \log n}{n-1} \approx \log n, \tag{5}$$

Lotka's law $g(1|k) \approx \frac{1}{k(k+1)}$, and Zipf's law $g(1||f) \approx \frac{1}{f}$.

What Davis (2018) did not show, we have

$$\begin{aligned}\mathbb{E}\, R_f \approx g(n||f) &= \frac{\log n - \sum_{j=1}^{f-1} (1-1/n)^j /j}{(1-1/n)^f} \\ &= \sum_{j=0}^{\infty} \frac{(1-1/n)^j}{j+f} \approx \exp\left(\frac{f}{n}\right) \Gamma\left(0, \frac{f}{n}\right).\end{aligned} \tag{6}$$

Model 3: Logistic model

The logistic model is the sigmoid hapax rate function of form

$$h(u) = \frac{1}{1 + e^u}. \tag{7}$$

This implies an asymptotically bounded vocabulary,

$$\mathbb{E}\, V \approx g(n) = \frac{2n}{n+1} \xrightarrow[n\to\infty]{} 2. \tag{8}$$

We have

$$\mathbb{E}\, R_f \approx g(n||f) = \frac{2n^f}{(n+1)^f}. \tag{9}$$

The inverse rank-frequency function decays like a geometric series!

Model 4: Linear model

We may be tempted to propose a piecewise linear hapax rate function,

$$h(u) = \begin{cases} 1, & u < 0, \\ 1 - \gamma u, & 0 \le u \le \gamma^{-1}, \\ 0, & u > \gamma^{-1}, \end{cases} \quad \gamma \approx 0.05. \tag{10}$$

This model is not an analytic function and it is ill-defined.
Nonetheless, the corresponding vocabulary size is

$$\mathbb{E}\, V \approx g(n) = \begin{cases} n, & n \le 1, \\ n^{1-\frac{1}{2}\gamma \log n}, & 1 \le n \le \exp(\gamma^{-1}), \\ \sqrt{\exp(\gamma^{-1})}, & n > \exp(\gamma^{-1}). \end{cases} \tag{11}$$

Function $\mathbb{E}\, R_f \approx g(n||f)$ follows from polynomials $h(u|k) = \frac{1}{k!} \sum_{j=0}^{k} a_{kj} u^j$, where we have the recursion

$$a_{kj} := \begin{cases} 0, & j < 0 \text{ or } j > k, \\ -1, & k = 0 \text{ and } j = 0, \\ \gamma a_{k-1,j-1} + (k-2) a_{k-1,j} - (j+1) a_{k-1,j+1}, & k \ge 1 \text{ and } 1 \le j \le k. \end{cases}$$

Model transformations

Suppose that we have some candidates for functions $h(u)$, $g(n)$, and $g(n||f)$.

These functions can be modified via:

- Offset: For an $\alpha \in \mathbb{R}$,

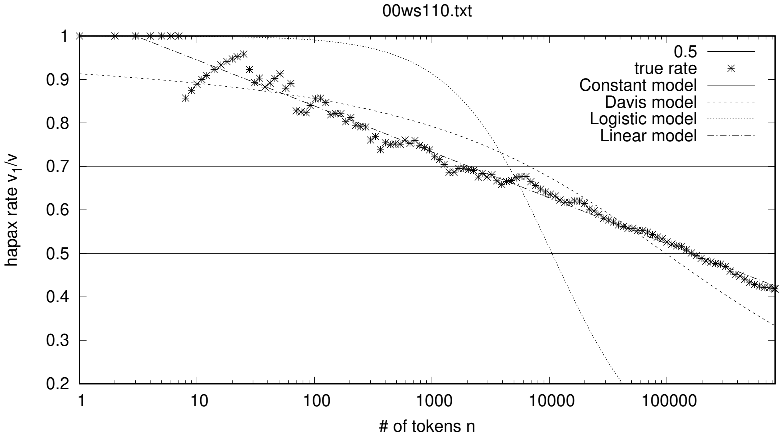
$$h_\alpha(u) := h(u-\alpha),$$
$$g_\alpha(n) := \frac{g(ne^{-\alpha})}{g(e^{-\alpha})},$$
$$g_\alpha(n||f) := \frac{g(ne^{-\alpha}||f)}{g(e^{-\alpha})}.$$

- Mixture: For a $\lambda \in (0,1)$,

$$h_\lambda(u) := \frac{\lambda h_1(u)g_1(e^u) + (1-\lambda)h_2(u)g_2(e^u)}{\lambda g_1(e^u) + (1-\lambda)g_2(e^u)},$$
$$g_\lambda(n) := \lambda g_1(n) + (1-\lambda)g_2(n),$$
$$g_\lambda(n||f) := \lambda g_1(n||f) + (1-\lambda)g_2(n||f).$$

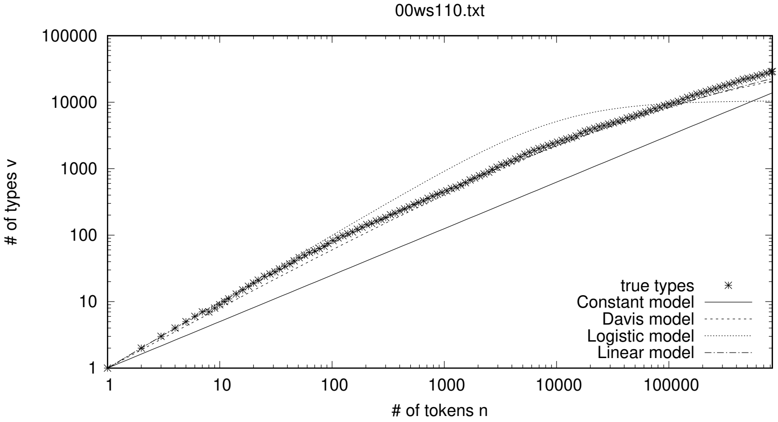
Predicted hapax rate

The hapax rate for Shakespeare's First Folio.



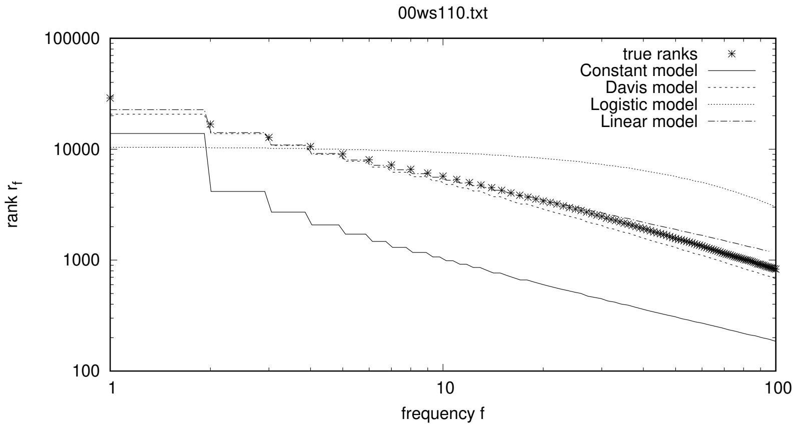
Predicted type-token plot

Below we present how the model fits to Shakespeare's First Folio.



Predicted rank-frequency plot

Below we present how the model fits to Shakespeare's First Folio.



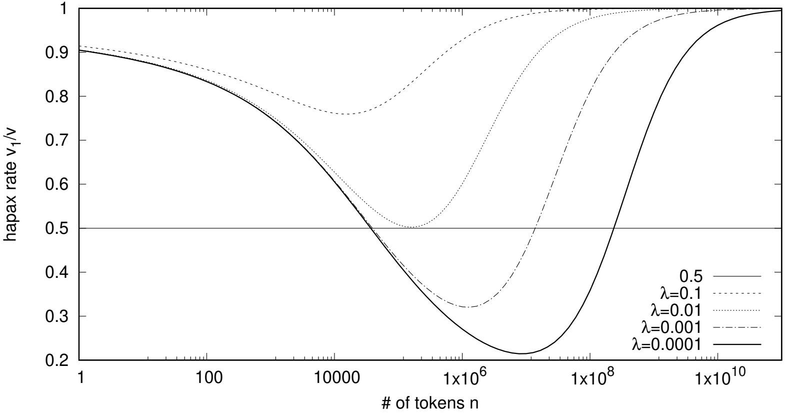
Fitted parameters — Project Gutenberg (English)

File	Constant	Davis	Logistic	Linear		Length
	β	α	α	γ	α	n
00ws110.txt	0.699	11.49	9.26	0.0459	1.101	835726
1ours10.txt	0.723	10.67	8.42	0.0491	1.073	128963
2000010.txt	0.732	10.37	8.32	0.0605	2.208	101247
2cahe10.txt	0.727	11.21	9.06	0.0495	1.624	298339
5wiab10.txt	0.754	10.92	8.45	0.0493	1.508	92558
800lg10.txt	0.653	9.43	8.16	0.051	0	95493
csnva10.txt	0.665	10.94	9.1	0.0508	1.229	1268149
dbrry10.txt	0.706	10.49	8.47	0.0494	0.846	159710
dscmn10.txt	0.639	9.62	8.66	0.052	0.201	312075
gltrv10.txt	0.716	10.2	8.26	0.0582	1.754	104909
milnd10.txt	0.701	10.18	8.35	0.062	2.112	195064
mt7bg10.txt	0.671	10.6	9.05	0.048	0.49	519886
stlla10.txt	0.681	10.29	8.31	0.0523	0.973	245882
wmcry10.txt	0.728	10.72	8.57	0.0532	1.666	145487
Mean	0.7	10.51	8.6	0.0522	1.199	321678

The second regime for large corpora

Fengxiang (2010) reported a $\boldsymbol{U}$-shaped plot for large corpora.

We can model it, for instance, with a mixture of the Davis model with $\alpha = 10.51$ and the constant model with $\beta = 1$:



Conclusion

Zipf's law plot with swapped axes is easier to analyze!

- It suffices to assume a simple analytic hapax rate function to derive the vocabulary size and the inverse rank-frequency function for any text size.
- These corrections to Zipf's and Herdan's laws contain one or two parameters but they apply to a wide range of text sizes.
- We plan a more extensive empirical verification.

Davis's model seems more precise than **Herdan's law**!

Still, we need better models of the hapax rate function!

References

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

V. Davis. Types, tokens, and hapaxes: A new heap's law. *Glottotheory*, 9(2):113–129, 2018.

F. Fengxiang. An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics*, 36(4): 631–637, 2010.

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

J. Milička. Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, 2(1):99–110, 2009.