On the Power Laws of Language: Word Frequency Distributions

Flavio Chierichetti*
Sapienza University
Rome, Italy
flavio@di.uniroma1.it

Ravi Kumar
Google
Mountain View, CA, USA
ravi.k53@gmail.com

Bo Pang
Google
Mountain View, CA, USA
bopang42@gmail.com

ABSTRACT

About eight decades ago, Zipf postulated that the word frequency distribution of languages is a *power law*, i.e., it is a straight line on a log-log plot. Over the years, this phenomenon has been documented and studied extensively. For many corpora, however, the empirical distribution barely resembles a power law: when plotted on a log-log scale, the distribution is concave and appears to be composed of two differently sloped straight lines joined by a smooth curve. A simple generative model is proposed to capture this phenomenon. The word frequency distributions produced by this model are shown to match the observations both analytically and empirically.

1 INTRODUCTION

The distribution of word frequencies is a fundamental phenotype of a language. Word frequency distributions have been studied by statisticians and linguists since the statistics of word usage yield valuable insights into the language, its construction, and its evolution. These distributions have been long-studied outside of statistics and linguistics as well. In information retrieval, word frequency distributions (and sometimes the ranks of word frequency) are directly used by many algorithms for many tasks, e.g., weighting the significance of documents and query terms [2, 36], text classification [6, 26], topic distillation [7, 13, 38], latent semantic analysis [24, 25], and so on. The word frequency distribution plays a central role in determining the size of inverted indices [14, 30], the compression ratio of natural texts [11, 12].

In his pioneering work, Zipf postulated that the frequency of any word in the language is inversely proportional to a power of its rank [44, 45]. On a log-log plot, with the x-axis representing the rank, and the y-axis representing the frequency, the distribution would thus appear as a straight line with a negative slope. Subsequent studies have confirmed similar phenomena on different corpora and genre. Even though the actual parameters can depend on the corpus, the power-law phenomenon itself was shown to be pervasive and robust. There have been many attempts to explain and refine Zipf's law [10, 15, 19, 20, 28, 29, 31, 34, 40, 41]. Additionally, Zipf's law has been considered in the context of document retrieval by IR systems [1–3, 8, 9].

In large-scale empirical studies, however, the rank-vs-frequency distributions do not appear as straight lines on a log-log plot. Instead, they exhibit a bend that makes the curve look concave; we call the rank value at which the bend occurs as the *knee*. Interestingly, the bend is consistent with Zipf's original plot: the maximum rank in his plots is close to 10^3, whereas the knee is usually observed at a rank that is an order of magnitude higher. It is likely that the lack of computing power and automated tools made it infeasible for Zipf to move to a rank significantly larger than 10^3. This concavity-in-the-tail phenomenon has been noted empirically [23].

In this work we focus on the concavity phenomenon of the word frequency distribution. We postulate that the concavity arises from a seamless fusion of two power laws around the knee; this fusion is the byproduct of a natural corpus generative model that we introduce. To validate our model, we examine a variety of corpora, ranging from novels to news articles, and fit the functional form that comes out of our process to their word frequency distributions. The fit is surprisingly accurate, at the head, the tail, and the knee of the distributions.

Informally stated, our model works in two stages. In the first stage, a vocabulary for the corpus is generated by choosing the words from a power law distribution on the language. In the second stage, the corpus is generated by sampling the vocabulary words according to the same or another power law distribution. We show that this two-stage process gives rise to a distribution that is made up of two fused power laws. We validate this model by showing that the distortion between the distributions produced by our model and the empirical distributions is quite small. We also argue that a double Pareto distribution, which is a natural candidate to explain two fused power laws, would not be able to produce such a small distortion. The use of a two-stage process is convenient for modeling corpora obtained from different topics (e.g., sports, politics), where the first stage selects the topic vocabulary. Latent factor models [24, 25] also use a two-level process for text generation; however, each word in the text is determined by a mixture of topics rather than a single topic.

We then turn our attention to the distribution of k-grams, which has also been studied [17, 21]. It has been observed for some English and Chinese corpora that the distribution becomes flatter as k increases [22, 23]. However, to the best of our knowledge, no work has tried to explain this phenomenon. We prove analytically that the k-gram distributions become flatter as k increases, under the simple assumption that the head of the word frequency distribution is a power law.

*Supported in part by the ERC Starting Grant DMAP 680153, by a Google Focused Research Award, and by the SIR Grant RBSI14Q743.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080821

2 RELATED WORK

Power laws, also known as Pareto distributions or Zipf's laws, have been observed in a broad range of settings, i.e., city populations, sizes of earthquakes, number of citations received by papers, sales of books, number of hits on webpages, etc. [33]. According to Mitzenmacher [32], "The first known attribution of the power law distribution of word frequencies appears to be due to Estoup [18], although generally the idea (and its elucidation) are attributed to Zipf [44–46]."

The power law, as stated by Zipf [46] ($y = Kx^{-\alpha}$), appears as a straight line on a log-log plot. Mandelbrot [29] extended this form to $y = K(B + x)^{-\alpha}$ to obtain a more accurate fit for high-frequency words. On the other hand, Simon [41] developed a stochastic process based on the work of Yule [43]. Sichel [40] studied empirical fit of word frequencies with compound Poisson distribution. Baayen [5] compared different statistical models proposed in previous work for word frequency distribution. These early papers used small-scale datasets (by today's standards) and therefore do not necessarily provide a good fit to the tail of large datasets. Ha et al. [23] noted that on large-scale datasets, the word frequency distribution clearly has a concavity when plotted on the log-log scale. That is, the curve bends away from the single straight line predicted by Zipf's law. This phenomenon was more pronounced when they looked at the distribution of Chinese characters. Ha et al. [23] did not attempt to explain the form of the curves. The double Pareto distribution, which approximates the concavity with two straight lines, has also been considered for approximating word frequency distributions [14]; the fit achieved by our model is significantly more accurate than the one achievable by double Pareto distributions (see Section 5.3). Baayen [4] studied similarity relations between words and word frequency distribution. He also noted that function words straighten out the head of the distribution and complex words straighten the tail. Samuelsson [37] related Zipf's law to Turing's local re-estimation formula and van Leijenhorst and van der Weide [42] related Zipf's law to Heap's law.

There has been some work on studying the distribution of k-grams as well. Ha et al. [22, 23] studied k-grams in English and Chinese and observed that the distribution became flatter as k increased. Character k-grams and k-tuple distributions were also studied in [17] and [21]. None of these works attempted to explain these phenomena; in our work, we analytically establish the form of the k-gram distributions and show these become flatter as k increases.

Various generative models have been proposed for producing power laws. Zipf [46] hypothesized that the power law is the result of the "principle of least effort"; this was re-examined later by Mandelbrot [29], Ferrer i Cancho and Sole [20], and Ferrer i Cancho et al. [19], who developed arguments for deriving power law distributions based on information-theoretic considerations. In another line of research, people have argued that preferential attachment can lead to power laws. The general argument can be traced back to Yule [43], and a generalization was proposed by Simon [41]. Perc [34] proposes a preferential attachment process for the evolution of a language, and uses it to derive the Zipf's law. Power laws can also be obtained through the "monkeys typing randomly" (or "not-so-randomly") processes [15, 28, 31]. None of these works attempted to explain the concavity in the tail of the word frequency distribution. Mitzenmacher [32] provides a good survey on the topic of power laws.

3 EMPIRICAL ANALYSIS

We first study if the concavity in word frequency distribution is pervasive and robust, i.e., does it exist over a broad range of datasets and does it exist even when we restrict the data to a specific genre or topic? A plausible hypothesis for observing the concavity could be that for a collection of text restricted to a given genre or a given topic, the distribution would be straighter; and mixing such distributions leads to the concavity. To test this hypothesis, we constructed different datasets that can be split by different criteria. What we observe is that the concavity exists for each sub-sample.

3.1 Datasets

We conducted our empirical studies over the following four datasets. The first is `Gutenberg`, which is a mixed-genre, multi-topic, multilingual, and multi-author corpus of electronic books that are in the public domain from the Gutenberg project. We use the average of the birth and death years of the author as the approximation for the publication year of the book. We took the subset of 16,797 books that were written in English and has a publication year between the 17th century and the 20th century, and grouped them into four disjoint time periods (by century). The vocabulary sizes range from 100K words to over 800K words, and corpus sizes from 10 million to over 500 million tokens. In addition, we can also sample this dataset by authors.

The second dataset is `News`, which is a large-scale collection of news articles on two topics, namely, sports and politics. The third dataset is `ANC` (American National Corpus), which is a collection of American English, with written texts of different genres and transcripts of spoken data produced post 1990. The fourth dataset is `Europarl`, which is a multilingual collection of European Parliament proceedings [27]. It includes semantically equivalent content in 21 European languages. The size of the text in each language ranges from 10 million to 50 million tokens. This allows us to examine the word frequency distribution in multiple languages without having to worry whether differences were due to differences in topics or genres.

All datasets went through the same preprocessing, where all punctuation marks were removed, and all remaining tokens lowercased. Table 1 shows the main statistics of each of these four datasets.

dataset	vocabulary size (# types)	corpus size (# tokens)
`Gutenberg` 19th	400,876	185M
`News` (politics)	256,758	31M
`ANC` (written)	115,806	8.6M
`Europarl`	87,554	56M

Table 1: Details of the four datasets

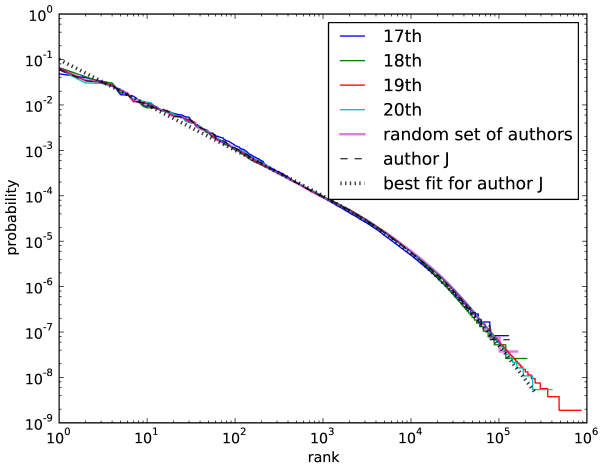

Figure 1: Word frequency distribution in the Gutenberg datasets: books in different time period (centuries), as well as a random subset of authors, and the subset of authors whose names begin with "J".

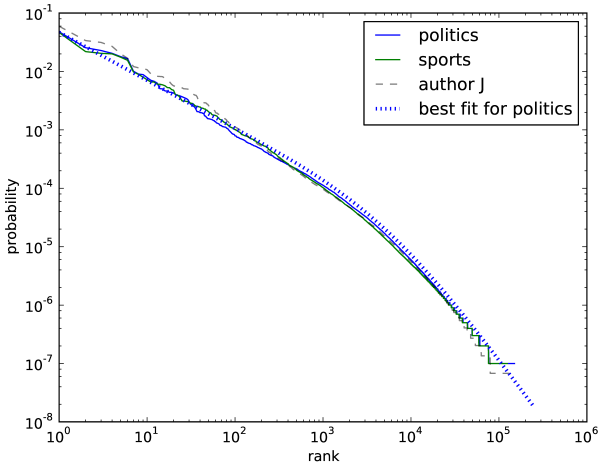

Figure 2: Word frequency distribution in news articles: politics vs sports

3.2 Word frequency distribution

First, we plot the empirical observations of word frequency distributions in different datasets on a log-log scale. We observe a clear concave shape over a broad range of corpora (Figures 1–3). Figure 1 shows word frequency distributions for different time periods in the `Gutenberg` dataset. As we can see, while the time periods (and vocabulary sizes) differ greatly, all curves closely resemble each other. In fact, the curve for `AuthorJ` (authors whose names begin with "J") largely follows the same shape. In subsequent plots, we include the `AuthorJ` curve as a reference point.

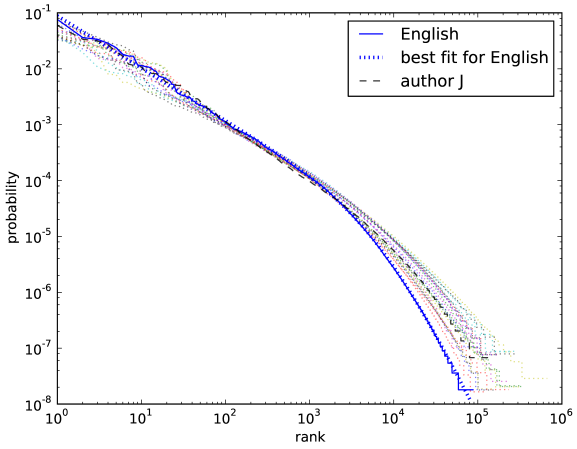

Figure 3: Word frequency distribution in 21 European languages. English is shown in blue solid line, other languages shown in dotted lines

Figure 2 shows word frequency distributions for two different topics (politics vs sports) in `News`. We observe a similar concave shape for both of them as `AuthorJ`. Figure 3 shows that the concave shape exists in a broad range of (21 European) languages. Furthermore, one could have hypothesized that smaller vocabulary leads to a straighter line (given previous studies with smaller datasets that focused on the straight-line Zipf distribution); but note the curve for English exhibits a more concave shape than that for `AuthorJ`, even though it is a smaller corpus and arguable over a more focused range of topics with less variations in styles.

3.3 k-gram frequency distribution

Figure 7 plots the empirical distribution of k-grams for $k = 1$ up to 5. Given the space constraints, we include only the plot for the `Europarl` data (where the unigram frequency distribution exhibits the highest degree of concavity). As observed in [22, 23] for some English and Chinese corpora, the lines get flatter as k grows.

4 MODEL

We define a simple and natural stochastic process for generating a corpus in a language. The process takes place in two stages. In the first stage, a *founding text* for the topic is written by choosing words from the language; the set of distinct words used in the founding text will form the *vocabulary* of the topic. In the second stage, the corpus itself is generated using the words in the vocabulary.

Let the parameters $\alpha, \beta \in (0, 1)$, $\gamma > 0$, and a positive integer n, be given. Here, α is the exponent of the power law $P^{(\alpha)}_{|U|}$ defined over the universe of words U; n will be length of the founding text; β determines the position of *knee* (which will be located around n^{β}). The parameter γ is not necessary, but lends more flexibility to our model as we will see below.

We set $N = \left\lceil n^{\frac{1-\alpha\beta}{1-\alpha}} \right\rceil = |U|$, i.e., N is the number of words in the language. The distribution on the language will be the power law

corpus	α	β	γ	n
Gutenberg 19th	0.618	0.795	1.034	400,876
News (politics)	0.569	0.725	0.898	256,758
ANC (written)	0.595	0.866	0.996	115,806
Europarl	0.691	0.800	0.986	87,554

Table 2: Fitted parameters for various corpora.

$P_N^{(\alpha)}$. As we will see later, this choice of N guarantees that the knee will be positioned close to the rank n^β.

(i) In the first stage, we choose n words independently from U according to the power law $P_N^{(\alpha)}$ to produce the founding text of the topic. The vocabulary V of the topic will be the set of words that appeared at least once in the founding text. As we will see below, with high probability we will have $|V| \approx n$.

(ii) In the second stage, we use a second power law $P_N^{(\gamma)}$ over the language, possibly but not necessarily different from $P_N^{(\alpha)}$. A *corpus* for the topic is generated by choosing words independently from $P_N^{(\gamma)}$ restricted to the vocabulary (i.e., support) V.

The use of a two-stage process is convenient to model corpora belonging to different topics (e.g., sports, politics): the first stage effectively determines the vocabulary of the topic. The assumption of choosing words from V according to the original power law $P_N^{(\alpha)}$ has been made before; see, for example [40].

In our model, the parameter γ gives additional flexibility, since one is not forced to use the same power law exponent α to choose words from the vocabulary. If we insist on model parsimony, the γ parameter can be removed by choosing γ as a function of α.

Figure 5 shows the best fit of our model (formalized in the next Section) to four empirical corpora. Note that the fit from our model traces the empirical distribution quite accurately (we discuss this in Section 5.4). We include the parameters for each fit in Table 2.

5 THEORETICAL ANALYSIS

We prove in this section that, while having a very small number of parameters, our model is able to generate curves that match the empirically observed curves. We then study the distribution of k-grams and prove that the distribution gets flatter as a function of k, which matches the empirical observation as well.

All the proofs missing from this section can be found in the Appendix.

5.1 Preliminaries

We begin with some basic notation. Let the upper incomplete *Gamma function* be given by $\Gamma(a,b) = \int_b^\infty x^{a-1}e^{-x}dx$; let $\Gamma(a) = \Gamma(a,0)$. The function $\Gamma(a,b)$ is well-defined for every real a if $b > 0$ and is well-defined for every $a > 0$ if $b = 0$. For each integer $k \geq 1$, we have $\Gamma(k) = (k-1)!$.

Let $\zeta(\alpha) = \sum_{i=1}^\infty i^{-\alpha}$ be the value of the *Riemann Zeta* function at $\alpha > 1$. For $\alpha > 0$, and an integer $N \geq 1$, let $\zeta_N(\alpha) = \sum_{i=1}^N i^{-\alpha}$.

Suppose that $\alpha > 1$, and let $P^\alpha(i) = i^{-\alpha}/\zeta(\alpha)$ for each integer $i \geq 1$. Then, $P^{(\alpha)}$ is the *power law distribution* with exponent α, defined on the universe $U = \{1, 2, \ldots\} = \mathbf{N}$.

Also, let $[N] = \{1, 2, \ldots, N\}$. For $\alpha > 0$, if we let $U = [N]$, we have that the truncated power law distribution on U is given by $P_N^{(\alpha)}(i) = i^{-\alpha}/\zeta_N(\alpha)$, for $i \in U$.

We first obtain some bounds on $\zeta_N(\alpha)$ and $P_N^{(\alpha)}(i)$.

Lemma 5.1. *For each $0 < \alpha < 1$, it holds that $\zeta_N(\alpha) = \frac{N^{1-\alpha}}{1-\alpha} \pm O(1)$ and $P_N^{(\alpha)}(i) = \left(1 \pm O\left(N^{\alpha-1}\right)\right) i^{-\alpha} \frac{1-\alpha}{N^{1-\alpha}}$.*

5.2 Word frequency distribution

In this section we analyze the word frequency distribution produced by the generative model. We proceed to study the probability $R(i)$ that the ith word, $1 \leq i \leq N$, appears in the vocabulary V. Since V was constructed using n independent samples, we have

$$R(i) = 1 - (1 - P_N^{(\alpha)}(i))^n.$$

By $N = \left(1 + O\left(n^{-\frac{1-\alpha\beta}{1-\alpha}}\right)\right) \cdot n^{\frac{1-\alpha\beta}{1-\alpha}}$, and by Lemma 5.1, we have

$$P_N^{(\alpha)}(i) = \frac{1-\alpha}{i^\alpha n^{1-\alpha\beta}} - O\left(i^{-\alpha} n^{2\alpha\beta-2}\right).$$

That is,

$$P_N^{(\alpha)}(i) = \left(1 - O\left(n^{\alpha\beta-1}\right)\right) \frac{1-\alpha}{i^\alpha n^{1-\alpha\beta}}.$$

Since $0 < \alpha, \beta < 1$ the multiplier is no worse than $1 - o(1)$.

Our analysis begins by showing that $R(i)$ — that is, the probability that the ith term of the language appears in the vocabulary — can be expressed by a simple exponential term.

Lemma 5.2. $R(i) = (1 \pm o(1))\left(1 - e^{-(1-\alpha)\frac{n^{\alpha\beta}}{i^\alpha}}\right)$.

Lemma 5.2 can be shown by approximating $P_N^{(\alpha)}(i)$ as in Lemma 5.1, since the error term in Lemma 5.1 is small enough to prove the statement of Lemma 5.2.

We then use the new expression of $R(i)$ to compute the expected number of terms with rank at most k that appears in the vocabulary. Specifically, let $U_k \subseteq U$ be the set of words that have rank at most k in $P_N^{(\alpha)}$. We focus on the number of words in U_k that make it to the vocabulary V. I.e., we focus on the random variable $|V \cap U_k|$. Observe that $E[|V \cap U_k|] = \sum_{i=1}^k R(i)$. Lemma 5.3 shows that this expectation is very well approximated by the function $D(k)$ (we use this notation as a shorthand for $D_{\alpha,n^\beta}(k)$):

$$D(k) = k - \frac{(1-\alpha)^{1/\alpha}}{\alpha} n^\beta \Gamma\left(-\frac{1}{\alpha}, (1-\alpha)\left(\frac{n^\beta}{k}\right)^\alpha\right).$$

Lemma 5.3. $E[|V \cap U_k|] = (1 \pm o(1))D(k)$.

The above lemma can be shown by integrating the expression of $R(i)$ that we obtained in Lemma 5.2, and by controlling the error term.

The next step of the proof is showing that $D(k)$ behaves like a simple polynomial in the ranges $k < o(n^\beta)$ and $k > \omega(n^\beta)$, i.e., for all k far enough from the knee. This will be key for proving that the head and the tail of the final distribution will be power laws.

Lemma 5.4. *If $k < o\left(n^\beta\right)$, then $D(k) = (1 \pm o(1))k$. If $k > \omega\left(n^\beta\right)$, then $D(k) = (1 \pm o(1))n^{\alpha\beta}k^{1-\alpha}$.*

The proof of Lemma 5.4 is relatively simple. We just have to use the approximations of the $R(i)$'s given by Lemma 5.2, i.e., $R(i) = 1 - o(1)$ if $i < o\left(n^\beta\right)$, and $R(i) = (1 \pm o(1))n^{\alpha\beta}\frac{1-\alpha}{i^\alpha}$ if $i > \omega\left(n^\beta\right)$. Then, the linearity of expectation, and Lemma 5.2, directly entail the claim.

Negative dependence can be used to prove the next Lemma, which simply states that with high probability for each $k \in [N]$, the random variable $|V \cap U_k|$ will be quite close to its expectation $E[|V \cap U_k|]$; by Lemma 5.3, that random variable will then be very close to the function $D(k)$ itself.

Lemma 5.5. *With probability* $1 - o(1)$*, we will have that for each* $k \in [N]$*,*

$$|V \cap U_k| = (1 \pm o(1)) \cdot D(k).$$

Lemma 5.5 allows us to get an expression for the frequency curve of the corpus. For $k \geq 1$, the frequency curve can be expressed parametrically as

$$x(k) = D(k), \quad \text{and} \quad y(k) = W \cdot k^{-\gamma},$$

where W is the normalization factor. In other words, the abscissa $x(y)$ that one has to associate to a given ordinate y is equal to $x(y) = D\left(\left(\frac{W}{y}\right)^{1/\gamma}\right)$.

Finally, we can state our main result about the word frequency distribution of the generative model. It follows as a corollary from Lemma 5.4 and Lemma 5.5.

Theorem 5.6. *With probability* $1 - o(1)$*, we will have:*
(i) $|V| = (1 \pm o(1))n$*, and*
(ii) for each rank $1 \leq k \leq |V|$*, the probability associated to the word of rank* k *in* V *will be proportional to*

$$\begin{cases} (1 \pm o(1)) \cdot k^{-\gamma}, & \text{if } k = o\left(n^\beta\right), \\ (1 \pm o(1)) \cdot \left(\frac{k}{n^{\alpha\beta}}\right)^{-\frac{\gamma}{1-\alpha}}, & \text{if } k = \omega\left(n^\beta\right). \end{cases}$$

Theorem 5.6 states the main properties of the word frequency distribution: the model produces a vocabulary of size close to n, the head of the vocabulary frequency curve follows a power law with exponent $-\gamma$, and the tail follows a power law with exponent $-\gamma/(1-\alpha)$. Moreover, our parametric definition of the curve gives a precise description of how the two power laws merge in one another. This is important for us since we want to precisely fit the curve to the datasets we have. Figure 4 shows the word frequency distribution produced by our model. Observe that our model produces two power laws that are joined around the knee at n^β, as expected.

5.3 Relation with double Pareto

One might wonder why we did not use a simple double Pareto curve (as in [14]) to model the distributions. The main reason is that the ratio of the probability at the rank $i \simeq n^\beta$ of the double Pareto curve (with the correct power laws, and the correct knee) and of our curve at the same i is quite large. We will show in this section that (i) the multiplicative distortion is at least $\left(\frac{e}{e-1}\right)^\gamma = (1.5819\ldots)^\gamma$ for $\alpha \to 0$, and (ii) the distortion diverges exponentially to ∞ as α approaches 1. Moreover, we numerically obtain that at $\alpha = 0.6$ (close to empirical numbers; see Table 2), the distortion becomes $(3.0270\ldots)^\gamma$. Later in Section 5.4, we empirically analyze the distortion in two of our corpora.

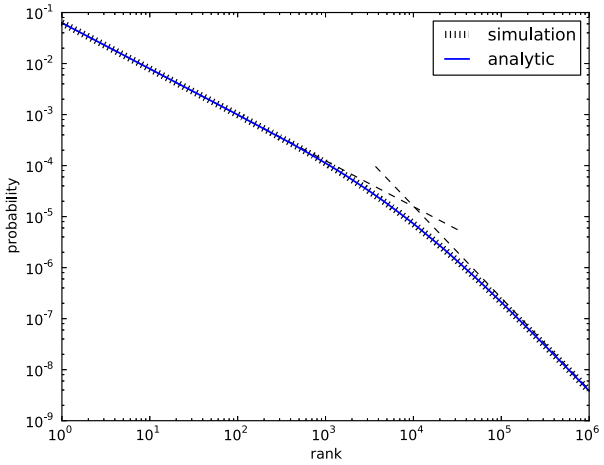

Figure 4: The result of an execution of the stochastic process with $\alpha = \frac{1}{2}, \beta = \frac{2}{3}, \gamma = \frac{9}{10}, n = 10^6$ and $N = n^{\frac{1-\alpha\beta}{1-\alpha}} = 10^8$. In this execution, the vocabulary V ended up with a cardinality of $|V| = 984328$, i.e., so many distinct words were randomly selected in the first stage of the process. The curve represents the probability distribution of the vocabulary. The head of the curve follows a power law with exponent $-\gamma = -0.9$ and the tail of the curve follows a power law with exponent $-\frac{\gamma}{1-\alpha} = -1.8$. Observe that the two power laws cross at an abscissa value close to $n^\beta = 10^4$.

We now show how this distortion can be computed, and obtain its limiting values at $\alpha = 0, 1$. First, for a given α, let k_α be the minimum integer such that $D(k_\alpha) \geq n^\beta$. Observe that $k_\alpha \geq n^\beta$. The probability of the $\lceil n^\beta \rceil$th word in our model's dictionary will then be $(1 \pm o(1)) \cdot W \cdot k_\alpha^{-\gamma}$ for some normalizing factor W.

Consider the double Pareto curve having the same head and tail power laws of our curve, and the same knee n^β. The value of this double Pareto curve at the $\lceil n^\beta \rceil$th word will be $(1 \pm o(1)) \cdot W \cdot n^{-\beta\gamma}$. Therefore, the distortion of the two curves at the $\lceil n^\beta \rceil$th word (i.e., at the knee) for a given α, as n tends to infinity, is at least $(d_\alpha)^\gamma$, where

$$d_\alpha = \liminf_{n\to\infty} \frac{k_\alpha}{n^\beta}.$$

We show in Lemma 5.7 that $\lim_{\alpha\to 0^+} d_\alpha = \frac{e}{e-1}$ Moreover, we will show in Lemma 5.8 that, as α converges to 1, d_α diverges at least as fast as $c^{1/(1-\alpha)}$ for some constant $c > 1$.

We state the two Lemmas, and briefly comment on how they can be proved.

Lemma 5.7. *For* $0 < \alpha < \frac{1}{2}$*, we have*

$$k_\alpha = \left(\frac{e}{e-1} \pm O\left(\alpha \ln 1/\alpha\right)\right) \cdot n^\beta.$$

A heuristic proof of the above statement would argue that, if $\alpha = 0$, then all the terms are equally likely to be chosen, i.e., they have probability N^{-1}, with $N = n^{\frac{1-\alpha\beta}{1-\alpha}} = n$. In other words, the

vocabulary is constructed by throwing n balls (the words in the founding text), into $N = n$ bins (the words of the language). By the Chernoff bound, any given set of $t \gg 1$ bins will be hit by approximately t balls with high probability. Moreover, by classic balls-in-bins arguments, the $\frac{e}{e-1} \cdot n^\beta$ balls that hit the first $\frac{e}{e-1} \cdot n^\beta$ bins will be distributed across approximately n^β distinct bins with high probability (essentially because of the Poissonian approximation of the binomial distribution). Hence, $k_\alpha \approx \frac{e}{e-1} \cdot n^\beta$.

The above reasoning can be made formal, so that it can be applied to small $\alpha > 0$.

Finally, we show that in the opposite regime, $\alpha = 1 - \epsilon$, d_α diverges to infinity at an exponential rate.

Lemma 5.8. *There exists a constant $c > 1$ such that, for all $\frac{1}{2} < \alpha < 1$, we have*

$$k_\alpha \geq c^{\frac{1}{1-\alpha}} \cdot n^\beta.$$

The above statement can be proven by partitioning the set of words of index up to $c^{\frac{1}{1-\alpha}} \cdot n^\beta$ into buckets in such a way that words in a given bucket have roughly the same probability of being selected in the vocabulary. The bucketing makes it easy to compute the expected number of words per bucket that end up in the vocabulary. Finally, adding up these expected numbers gives the above lower bound.

5.4 Fitting

The parameters of the fitting were obtained by minimizing the Kullback-Leibler divergence $D_{\mathrm{KL}}(P||E)$ of our model's frequency distribution P from the empirical distribution E. I.e., given E, we searched for the P that minimizes $D_{\mathrm{KL}}(P||E) = \sum_i \left(P(i) \cdot \ln \frac{P(i)}{E(i)}\right)$.

More precisely, if the empirical distribution E was over n distinct words then, given a triple of parameters (α, β, γ), we computed a candidate distribution $P = P_{n,\alpha,\beta,\gamma}$ by letting, for each $i = 1, \ldots, n$, $P(i)$ be proportional to $k_i^{-\gamma}$ with k_i equal to the solution of $i = D_{\alpha,n^\beta}(k_i)$.

We used a brute-force approach to guess the optimal fitting parameters α, β, γ. The results are reported in Table 2. The empirical curves, and their fittings, are shown in Figure 5. To show how much better our curve's fits are (with respect to the double Pareto fits), we plot in Figure 6 the ratios between the probabilities given by our curve and the actual distribution, and those given by double Pareto and the actual distribution, for the `Gutenberg` and `News` corpora.

6 K-GRAM FREQUENCY DISTRIBUTION

Let $P^{(\alpha)}$ be the power law distribution with exponent $\alpha > 1$ over an infinite language $U = \mathbb{N}$. Given an integer $k \geq 1$, let $P^{(\alpha,k)}$ be the probability distribution on k-tuples $\langle u_1, \ldots, u_k \rangle$, where $u_1, \ldots, u_k \in U$ are chosen independently from $P^{(\alpha)}$. We now show analytically that (i) the distribution of $P^{(\alpha,k)}$ will be close to the original power law $P^{(\alpha)}$ and (ii) the curves corresponding to k-grams will become flatter as k increases, when plotted on a log-log scale; this phenomenon can be observed empirically in Figure 7.

Theorem 6.1. *If we sort the k-tuples decreasingly by their probabilities in $P^{(\alpha,k)}$, then the probability of the rth k-tuple will be equal*

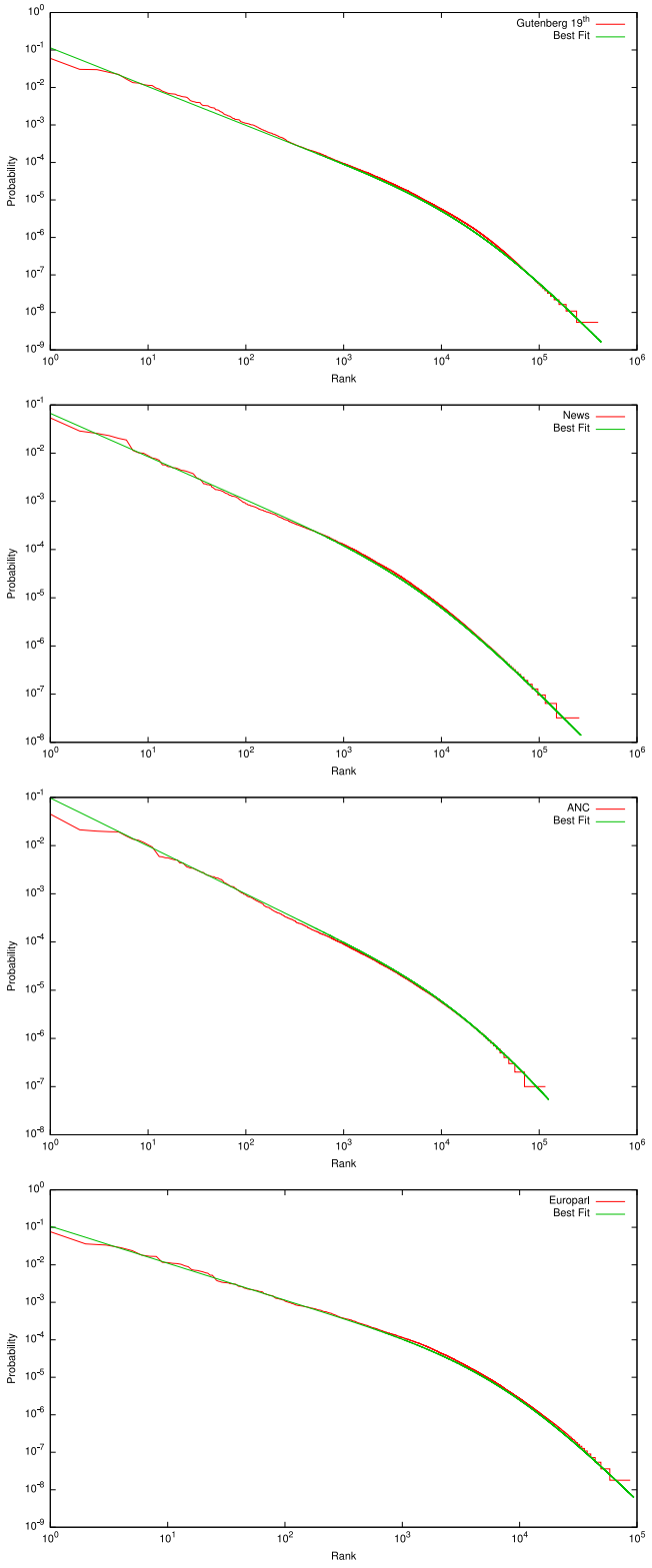

Figure 5: The empirical and the (fitted) theoretical curves of four corpora.

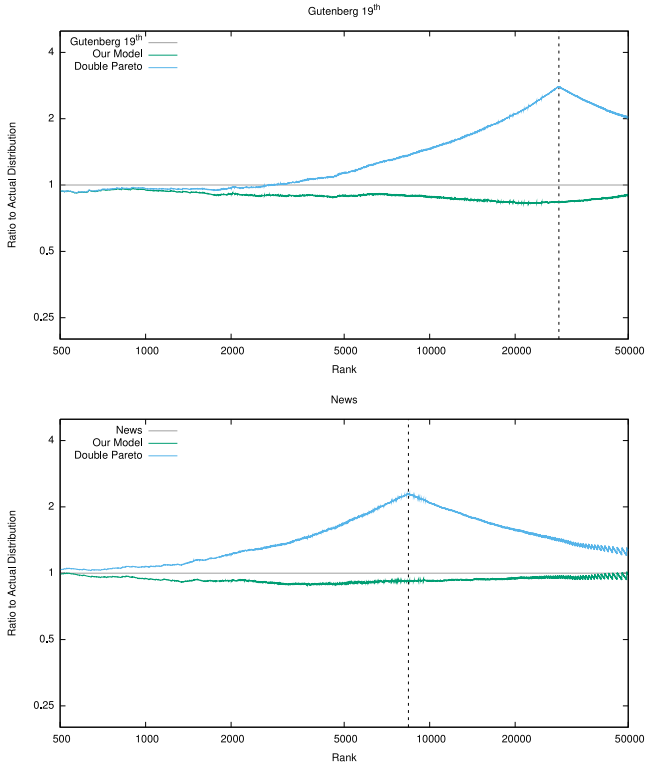

Figure 6: A log-log plot of the ratios between our fitted curve (resp., the Double Pareto curve) and the empirical curve, at word ranks 500 to 50000 (that is, around the knee). The vertical dashed line represents the position of the Double Pareto knee. Our curve is a very good multiplicative approximation of the empirical curve (the ratios induced by our curve are quite close to 1), and generally a much better approximation than Double Pareto; the maximum ratios, or distortions (see Section 5.3), incurred by Double Pareto happen around the knees: the ratios, there, are close to 2.8 in the `Gutenberg` corpus, and to 2.3 in the `News` corpus.

to

$$(1 \pm o_r(1)) \cdot \zeta(\alpha)^{-k} \cdot \Gamma(k)^{-\alpha} \cdot \left(\frac{\ln^{k-1} r}{r}\right)^{\alpha}.$$

Proof. Our goal is to compute the position (or rank) $r_{i_1,\ldots,i_k}$ of the product $P^{(\alpha)}(i_1)\cdots P^{(\alpha)}(i_k)$ in the ordered multiset

$$\left\{P^{(\alpha)}(j_1)\cdots P^{(\alpha)}(j_k) \mid j_1,\ldots,j_k \in \mathbf{Z}^+\right\}.$$

In other words, we aim to compute the number of tuples $\langle j_1,\ldots,j_k\rangle$ such that $P^{(\alpha)}(i_1)\cdots P^{(\alpha)}(i_k) < P^{(\alpha)}(j_1)\cdots P^{(\alpha)}(j_k)$. Rewriting this condition using the fact $p(i) \propto i^{-\alpha}$ and letting $n = i_1\cdots i_k$, we get

$$\begin{aligned} r = r_n &= |\{\langle j_1,\ldots,j_k\rangle \mid j_1\cdots j_k < n\}| \\ &= \frac{n\ln^{k-1} n}{\Gamma(k)} + O(n\ln^{k-2} n), \end{aligned}$$

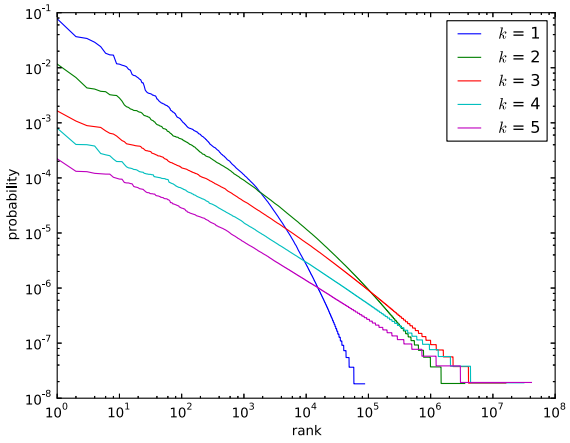

Figure 7: k-grams frequency distribution for the Europarl dataset (English).

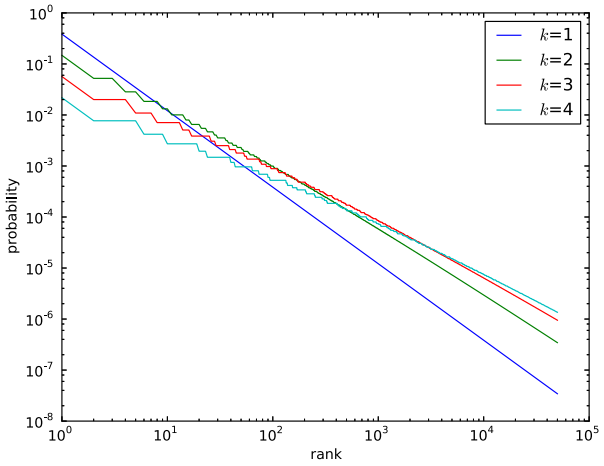

Figure 8: Computed k-gram distribution, $\alpha = 1.5$.

which follows from [35, 39]. Inverting this, we obtain

$$n = (1 + o_r(1))\Gamma(k)\frac{r}{\ln^{k-1} r}.$$

The proof is concluded by recalling that the probability of the tuple $\langle i_1,\ldots,i_k\rangle$ is

$$P^{(\alpha)}(i_1)\cdots P^{(\alpha)}(i_k) = \zeta(\alpha)^{-k} n^{-\alpha}. \quad \square$$

Observe that our Theorem gives sharp bounds on the probability of the rth k-gram, as r diverges. Figure 8 shows the k-gram distribution computed with a synthetic power law distribution with $\alpha = 1.5$; Figure 9 shows the curves predicted by Theorem 6.1. We can see that the empirical curves agree asymptotically with the theoretical estimates.

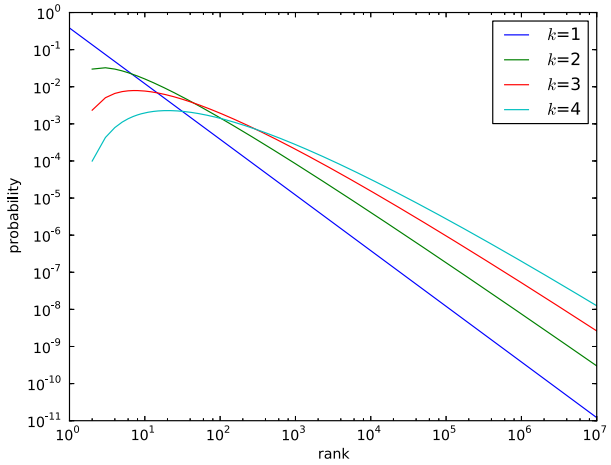

Figure 9: k-grams distribution using Theorem 6.1, $\alpha = 1.5$.

7 CONCLUSIONS

In this paper we took a closer look at the word frequency distribution. We observed a knee, leading to a concavity, in the empirical distributions of many different kinds of corpora, and proposed a natural text generation model to explain the knee and the concavity. We then analytically showed that our model produces distributions nearly identical to the empirically observed ones. We also analyzed the k-gram distribution that one obtains by picking words independently from a power law distribution. We proved that the k-gram distribution becomes flatter as k increases; this phenomenon had only been empirically observed in the literature but never analyzed. Our generative model opens up many interesting questions: can the distributions it produces be used in applications such as text compression, translation, and information retrieval?

REFERENCES

[1] IJsbrand Jan Aalbersberg. 1991. Posting compression in dynamic retrieval environments. In *SIGIR*. 72–81.
[2] IJsbrand Jan Aalbersberg. 1994. A document retrieval model based on term frequency ranks. In *SIGIR*. 163–172.
[3] Leif Azzopardi. 2009. Query Side Evaluation: An empirical analysis of effectiveness and effort. In *SIGIR*. 556–563.
[4] Harald Baayen. 1991. A stochastic process for word frequency distributions. In *ACL*. 271–278.
[5] Harald Baayen. 1992. Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities* 26, 5 (1992), 347–363.
[6] L. Douglas Baker and Andrew Kachites McCallum. 1998. Distributional clustering of words for text classification. In *SIGIR*. 96–103.
[7] Krishna Bharat and Monika R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR*. 104–111.
[8] David C Blair. 1990. *Language and Representation in Information Retrieval*. Elsevier Science Publishers.
[9] David C. Blair. 2002. The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing & Management* 38, 2 (2002), 273–291.
[10] Andrew D Booth. 1967. A "Law" of occurrences for words of low frequency. *Information and Control* 10, 4 (1967), 386–393.
[11] Nieves R. Brisaboa, Antonio Fariña, Susana Ladra, and Gonzalo Navarro. 2008. Reorganizing compressed text. In *SIGIR*. 139–146.
[12] Nieves R. Brisaboa, Antonio Fariña, Gonzalo Navarro, and José R. Paramá. 2007. Lightweight Natural Language Text Compression. *Information Retrieval* 10, 1 (2007), 1–33.
[13] Soumen Chakrabarti, Mukul Joshi, and Vivek Tawde. 2001. Enhanced topic distillation using text, markup tags, and hyperlinks. In *SIGIR*. 208–216.
[14] Flavio Chierichetti, Ravi Kumar, and Prabhakar Raghavan. 2009. Compressed web indexes. In *WWW*. 451–460.
[15] Brian Conrad and Michael Mitzenmacher. 2004. Power laws for monkeys typing randomly: The case of unequal probabilities. *IEEE Transactions on Information Theory* 50, 7 (2004), 1403–1414.
[16] Devdatt Dubhashi and Alessandro Panconesi. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press.
[17] Leo Egghe. 2004. The distribution of N-grams. *Scientometrics* 47, 2 (2004), 237–252.
[18] Jean-Baptiste Estoup. 1916. *Gammes Sténographiques*. Institut Stenographique de France.
[19] Ramon Ferrer i Cancho, Oliver Riordan, and Béla Bollobás. 2005. The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society B: Biological Sciences* 272, 1562 (2005), 561–565.
[20] Ramon Ferrer i Cancho and Ricard V. Sole. 2003. Least effort and the origins of scaling in human language. *PNAS* 100, 3 (2003), 788–791.
[21] Xiaocong Gan, Dahui Wang, and Zhangang Han. 2009. *N-tuple Zipf analysis and modeling for language, computer program and DNA*. Technical Report 0908.0500v1. arXiv.
[22] Le Q Ha, P Hanna, D W Stewart, and F J Smith. 2006. Reduced n-gram models for English and Chinese corpora. In *COLING-ACL*. 309–315.
[23] Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. 2002. Extension of Zipf's law to words and phrases. In *COLING*. 1–6.
[24] Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *UAI*. 289–296.
[25] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *SIGIR*. 50–57.
[26] Thorsten Joachims. 2001. A statistical learning model of text classification for support vector machines. In *SIGIR*. 128–136.
[27] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. MT summit. (2005).
[28] Andreas Krause and Andreas Zollmann. 2002. Not so randomly typing monkeys–Rank-frequency behavior of natural and artificial languages. Algorithms for Information Networks–Project Report. (2002).
[29] Benoit Mandelbrot. 1953. An informational theory of the statistical structure of language. In *Communication Theory*, W. Jackson (Ed.). Butterworths, London, 486–502.
[30] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York.
[31] George A Miller. 1957. Some effects of intermittent silence. *The American Journal of Psychology* 70, 2 (1957), 311–314.
[32] Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1, 2 (2004), 226–251.
[33] Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 5 (2005), 323–351.
[34] Matja Perc. 2012. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface* 9, 77 (2012), 3323–33238.
[35] Adolf Piltz. 1881. *Über das Gesetz, nach welchem die mittlere Darstellbarkeit der natürlichen Zahlen als Produkte einer gegebenen Anzahl Faktoren mit der Grösse der Zahlen wächst*. Ph.D. Dissertation. University of Berlin.
[36] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523.
[37] Christer Samuelsson. 1996. Relating Turing's Formula and Zipf's Law. In *Proceedings of the 4th Workshop on Very Large Corpora*. 70–78.
[38] Issei Sato and Hiroshi Nakagawa. 2010. Topic Models with power-law using Pitman–Yor process. In *KDD*. 673–682.
[39] Atle Selberg. 1954. Note on a paper by L. G. Sathe. *J. Indian Math. Soc., N. Ser.* 18 (1954), 83–87.
[40] Herbert S Sichel. 1975. On a distribution law for word frequencies. *J. Amer. Statist. Assoc.* 70, 351a (1975), 542–547.
[41] Herbert A Simon. 1955. On a class of skew distribution functions. *Biometrika* 42, 3/4 (1955), 425–440.
[42] D.C. van Leijenhorst and Th.P. van der Weide. 2005. A formal derivation of Heaps' law. *Information Sciences* 170 (2005), 263–272.
[43] G Udny Yule. 1925. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 213 (1925), 21–87.
[44] George K. Zipf. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.
[45] George K. Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin Company.
[46] George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

APPENDIX

Proof of Lemma 5.1

First we upper bound $\zeta_N(\alpha)$ with one (the first term of its sum) plus the area under $x^{-\alpha}$ in the interval $[1, N]$,

$$\zeta_N(\alpha) \leq 1 + \int_1^N x^{-\alpha}\, dx = \frac{N^{1-\alpha} - \alpha}{1-\alpha}.$$

Analogously, a lower bound is given by the area under $(x+1)^{-\alpha}$ in the interval $[0, N]$,

$$\zeta_N(\alpha) \geq \int_0^N (x+1)^{-\alpha}\, dx \geq \frac{N^{1-\alpha} - 1}{1-\alpha}.$$

If $\alpha < 1$, using the expression for $P_N^{(\alpha)}(i)$, we get

$$i^{-\alpha}\frac{1-\alpha}{N^{1-\alpha}-\alpha} \leq P_N^{(\alpha)}(i) \leq i^{-\alpha}\frac{1-\alpha}{N^{1-\alpha}-1}. \quad \square$$

Proof of Lemma 5.2

If $i < o(n^\beta)$, then observe that, by $R(i) = 1 - (1 - P_N^{(\alpha)}(i))^n$ and $P_N^{(\alpha)}(i) = \left(1 - O\left(n^{\alpha\beta-1}\right)\right)\frac{1-\alpha}{i^\alpha n^{1-\alpha\beta}}$, we have

$$R(i) = 1 - o(1). \tag{1}$$

The right-hand expression in our claim simplifies to:

$$(1 \pm o(1))\left(1 - e^{-\omega(1)}\right) = 1 \pm o(1).$$

The claim is thus proved.

Next assume that i is a positive integer such that $i > \omega(n^{2\beta-1/\alpha})$. Observe that, by $\alpha, \beta < 1$, this case includes all the i's that are not part of the previous case.

For $0 < a < \frac{1}{2}$, and $b > 0$, it holds that

$$e^{-ab} \geq (1-a)^b \geq e^{-ab-2a^2b}.$$

Recall that $1 - R(i) = (1 - P_N^{(\alpha)}(i))^n$. Then, we bound the following quantities, using (5.2):

(1) $nP_N^{(\alpha)}(i) = (1-\alpha)n^{\alpha\beta}i^{-\alpha} + O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right)$. Observe that the error term is $o(1)$ for each i in our range.

(2) $2n\left(P_N^{(\alpha)}(i)\right)^2 = \Theta\left(\left(nP_N^{(\alpha)}(i)\right)P_N^{(\alpha)}(i)\right) = O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right)$, since $P_N^{(\alpha)}(i) = O\left(n^{\alpha\beta-1}\right)$ for each $i \geq 1$.

It follows that

$$1 - R(i) = e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha} \pm O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right)} = \left(1 \pm O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right)\right)e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}}.$$

Moreover, since $0 \leq e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}} \leq 1$, we have

$$R(i) = 1 - e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}} \pm O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right)$$

Observe that, if $i \leq n^\beta$, then we have $1 - e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}} \geq 1 - e^{-1+\alpha} = \Theta(1)$, while $O\left(\frac{n^{\alpha\beta}}{i^\alpha}n^{\alpha\beta-1}\right) \leq O\left(n^{\alpha\beta-1}\right)$. That is, $R(i) = (1 \pm o(1))\left(1 - e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}}\right)$.

On the other hand, if $i \geq n^\beta$, then if we let x be the exponent of the exponential term, we have $0 \leq x \leq 1$. For this range of x's, it holds $e^{-x} \leq 1 - \frac{x}{2}$ — equivalently, $1 - e^{-x} \geq \frac{x}{2}$. Therefore, $1 - e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}} \geq \frac{1}{2}(1-\alpha)n^{\alpha\beta}i^{-\alpha}$. Therefore, even for $i \geq n^\beta$, we have

$$R(i) = (1 \pm o(1))\left(1 - e^{-(1-\alpha)n^{\alpha\beta}i^{-\alpha}}\right). \quad \square$$

Proof of Lemma 5.3

For each integer $k \geq 1$ and for each non-decreasing and non-negative function $f(x)$ admitting a finite integral in $[0, k+1]$, we have

$$F_L = \int_0^k f(x)dx \leq \sum_{i=1}^k f(i) \leq \int_1^{k+1} f(x)dx = F_U.$$

Suppose that $0 \leq f(x) \leq 1$. Then,

$$F_U - F_L = \int_k^{k+1} f(x)dx - \int_0^1 f(x)dx \leq 1,$$

and hence $\sum_{i=1}^k f(i) = F_L + \xi$ for some $\xi \in [0, 1]$.

Observe that for all $q > 0$ and $\alpha \in (0,1)$, the function $f(x) = e^{-qx^{-\alpha}}$ satisfies the above conditions. We also have

$$\int f(x)dx = \frac{1}{\alpha}q^{1/\alpha}\Gamma\left(-\frac{1}{\alpha}, qx^{-\alpha}\right) + c,$$

where c is a constant. By choosing $q = (1-\alpha)n^{\alpha\beta}$, we get $e^{-(1-\alpha)\frac{n^{\alpha\beta}}{i^\alpha}} = f(i)$. Now,

$$F_L = \sum_{i=1}^k e^{-(1-\alpha)\frac{n^{\alpha\beta}}{i^\alpha}} - \xi = \lim_{\epsilon\to 0^+}\left[\frac{\left((1-\alpha)n^{\alpha\beta}\right)^{1/\alpha}}{\alpha}\Gamma\left(-\frac{1}{\alpha}, \frac{(1-\alpha)n^{\alpha\beta}}{x^\alpha}\right)\right]_{x=\epsilon}^k = n^\beta\frac{(1-\alpha)^{1/\alpha}}{\alpha}\Gamma\left(-\frac{1}{\alpha}, (1-\alpha)\cdot\left(\frac{n^\beta}{k}\right)^\alpha\right),$$

since by definition $\lim_{x\to\infty}\Gamma(a,x) = 0$. Thus, we have

$$\sum_{i=1}^k\left(1 - e^{-(1-\alpha)\frac{n^{\alpha\beta}}{i^\alpha}}\right) = D(k) - \xi. \tag{2}$$

Now, consider the first $o(n^\beta)$ terms of the LHS sum in (2). The exponent of e in each of them is $\omega(1)$, and therefore each of them has value $1 - o(1)$. It follows that the LHS of (2) has value $\omega(1)$. Since $0 \leq \xi \leq 1$, we have

$$(1 \pm o(1))\sum_{i=1}^k\left(1 - e^{-(1-\alpha)\frac{n^{\alpha\beta}}{i^\alpha}}\right) = D(k).$$

The claim then follows from Lemma 5.2 and the linearity of expectation. $\square$

Proof of Lemma 5.4

The first part is implied directly by (1) and Lemma 5.3. Hence, let $k = \omega\left(n^\beta\right)$. Let us define $g = \left\lceil\sqrt{n^\beta k}\right\rceil$ to be the ceiling of the geometric mean of n^β and k. Observe that $\omega\left(n^\beta\right) < g < o(k)$.

Since $R(i) \leq 1$, we have

$$\sum_{i=g+1}^{k} R(i) \leq E[|V \cap U_k|] \leq g + \sum_{i=g+1}^{k} R(i).$$

By Lemma 5.2, we have that $R(i) = (1 \pm o(1)) \cdot n^{\alpha\beta} \frac{1-\alpha}{i^{\alpha}}$ whenever $i > g$, since $g > \omega\left(n^{\beta}\right)$. Then, we can write:

$$\begin{aligned}\sum_{i=g+1}^{k} R(i) &= (1 \pm o(1))(1-\alpha)n^{\alpha\beta} \sum_{i=g+1}^{k} i^{-\alpha} \\ &= (1 \pm o(1))(1-\alpha)n^{\alpha\beta}\left(\zeta_k(\alpha) - \zeta_g(\alpha)\right) \\ &= (1 \pm o(1))n^{\alpha\beta}k^{1-\alpha},\end{aligned}$$

where the last step follows from Lemma 5.1. The value of the sum is $\Theta\left(n^{\alpha\beta}k^{1-\alpha}\right) = \omega\left(n^{\beta}\right)$, since $k > \omega\left(n^{\beta}\right)$. Therefore,

$$E[|V \cap U_k|] = (1 \pm o(1)) \cdot n^{\alpha\beta}k^{1-\alpha}.$$

Lemma 5.3 completes the proof. □

Proof of Lemma 5.5

Observe that, by Lemma 5.3, it is sufficient to prove that, with probability $1-o(1)$, it will happen that, for each $k \in [N]$, $|V \cap U_k| = (1 \pm o(1))E[|V \cap U_k|]$.

Let us define $X = n^{\beta}\log^{-\frac{1}{\alpha}} n$. We will use two arguments for proving the claim: one that holds if $k < o(X)$, and one that holds if $k > \omega\left(X^{1-\epsilon}\right)$, for any constant $0 < \epsilon < \frac{\beta}{4}$.

First, consider $k < o(X)$. Recall that we have $P_N^{(\alpha)}(i) = \Theta\left(\frac{1}{n}\left(\frac{n^{\beta}}{i}\right)^{\alpha}\right)$. Therefore, for $i < o(X)$, we have $P_N^{(\alpha)}(i) > \omega\left(\frac{\log n}{n}\right)$. For the same i's, therefore, we have

$$R(i) \geq 1 - \left(1 - \omega\left(\frac{\log n}{n}\right)\right)^{n} \geq 1 - n^{-\omega(1)}.$$

By the union bound, the probability that at least one of the terms of rank $i < o(X)$ in U does not end up in V is $n^{-\omega(1)}$. The claim is then proved for each $k < o(X)$.

Now consider $k > \omega\left(X^{1-\epsilon}\right)$. Let $Y_{i,j}$ be the indicator random variable of the event "the jth term of the founding text happened to be the ith term of the language". Then, for each j, the variables $Y_{1,j}, Y_{2,j}, \ldots, Y_{N,j}$ are negatively associated (see Chapter 3 of [16]). Moreover, by closure under product, the variables $Y_{i,j}$, for each $i \in [N], j \in [n]$ are as a whole negatively associated. Finally, since max is a monotone non-decreasing function the variables $Y_i = \max_{j=1,\ldots,n} Y_{i,j}$, $i \in [N]$, are also negatively associated — the Chernoff bound can then be applied to their sum, that is, to $\sum_{i=1}^{k} Y_i = |V \cap U_k|$. Thus, for each $0 < \delta < 1$,

$$\begin{aligned}&\Pr\left[\left||V \cap U_k| - E\left[|V \cap U_k|\right]\right| \geq \delta E\left[|V \cap U_k|\right]\right] \\ \leq &2e^{-\frac{\delta^2}{3}E[|V \cap U_k|]}.\end{aligned}$$

Since $k > \omega\left(X^{1-\epsilon}\right)$, we have $E[|V \cap U_k|] > \omega\left(n^{\beta-2\epsilon}\right)$. If we choose $\delta = n^{-\frac{1}{2}\beta+2\epsilon}$, we get:

$$\Pr\left[|V \cap U_k| = (1 \pm 2\delta) \cdot E\left[|V \cap U_k|\right]\right] \leq e^{-\Omega(n^{2\epsilon})}.$$

By the union bound, the claim is then proved for each $k > \omega\left(X^{1-\epsilon}\right)$. □.

Proof of Lemma 5.7

Suppose that $i = x \cdot n^{\beta}$. Then,

$$R(i) = (1 \pm o(1))\left(1 - e^{-(1-\alpha)x^{-\alpha}}\right).$$

Observe that, if $x \geq \alpha$, then $x^{-\alpha} \leq \alpha^{-\alpha} = e^{\alpha \ln \frac{1}{\alpha}} = 1 + O\left(\alpha \ln \frac{1}{\alpha}\right)$. Moreover, if $x \leq e$ then, $x^{-\alpha} \geq e^{-\alpha} = 1 - O(\alpha)$.

By the monotonicity of $x^{-\alpha}$ we thus obtain that $x^{-\alpha} = 1 \pm O\left(\alpha \ln \frac{1}{\alpha}\right)$ for all $x \in [\alpha, e]$.

For all $\alpha n^{\beta} \leq i \leq e n^{\beta}$, we then have

$$R(i) = (1 \pm O(\alpha \ln 1/\alpha)) \cdot \left(1 - e^{-1}\right).$$

For $i < \alpha n^{\beta}$, we have $R(i) \leq 1$. Therefore, for $\sum_{i=1}^{k_\alpha} R(i)$ to be at least n^{β}, we need

$$\frac{k_\alpha}{n^{\beta}} \geq \frac{1}{1-e^{-1}} - O(\alpha \ln 1/\alpha).$$

Moreover, for the inequality to hold, it suffices to have

$$\frac{k_\alpha}{n^{\beta}} \leq \frac{1}{1-e^{-1}} + O(\alpha \ln 1/\alpha). \quad \square$$

Proof of Lemma 5.8

Let us define $\epsilon = 1 - \alpha$. Recall that $E[|V \cap U_k|] = \sum_{i=1}^{k} R(i)$. Let $1 = t_0 < t_1 < \ldots < t_r = k$ be integers and, for $0 \leq j \leq r-1$, let p_j be any real number such that $p_j \geq R(t_j)$. Then, by the monotonicity of the $R(i)$'s, we have $p_j \geq R(i)$ for $1 \leq i \leq t_j$. Therefore,

$$\begin{aligned}E[|V \cap U_k|] &= \sum_{j=1}^{r} \sum_{i=t_{j-1}}^{t_j} R(i) \leq \sum_{j=1}^{r} \sum_{i=t_{j-1}}^{t_j} R(t_{j-1}) \\ &\leq \sum_{j=1}^{r} \left(t_j \cdot p_{j-1}\right).\end{aligned}$$

We set $t_0 = 1$ and, for $j \geq 1$, let $t_j = \left\lceil 2^{\frac{j-2}{\alpha}} \cdot n^{\beta} \right\rceil$. We let r be unspecified for now. Also, let $p_0 = 1$, and, for $j \geq 1$,

$$p_j = 1 - e^{-(1-\alpha)n^{\alpha\beta}t_j^{-\alpha}} = 1 - e^{-\epsilon 2^{2-j}} \leq O\left(\epsilon 2^{-j}\right).$$

We have that $\frac{1}{\alpha} - 1 = \frac{1}{1-\epsilon} - 1 \leq O(\epsilon)$. For $j = 1$, we have $t_j p_{j-1} = t_1 \leq \left\lceil \frac{n^{\beta}}{2} \right\rceil$. Moreover, for $j \geq 2$,

$$t_j \cdot p_{j-1} \leq O\left(\epsilon n^{\beta} 2^{j\left(\frac{1}{\alpha}-1\right)}\right) = \epsilon n^{\beta} 2^{O(j\epsilon)}.$$

As $j \leq O(1/\epsilon)$, the latter is at most $O(\epsilon n^{\beta})$. In fact, there exists a constant $b > 0$ such that if we let $r = \lceil b/\epsilon \rceil$, we have

$$E[|V \cap U_k|] \leq \left\lceil \frac{n^{\beta}}{2} \right\rceil + \sum_{j=2}^{r} O\left(\epsilon n^{\beta}\right) < \frac{3}{4} \cdot n^{\beta}.$$

Lemma 5.3 shows that $D(k) = (1 \pm o(1))E[|V \cap U_k|]$. With our choice of r, k equals

$$k = t_r = \left\lceil 2^{\frac{r-2}{\alpha}} \cdot n^{\beta} \right\rceil \geq c^{1/\epsilon} \cdot n^{\beta},$$

for some constant $c > 1$.

Therefore, $D\left(c^{1/\epsilon} \cdot n^{\beta}\right) \leq (1 \pm o(1))\frac{3}{4}n^{\beta}$. By the latter, and by the monotonicity in k of $E[|V \cap U_k|]$, we have that $k_\alpha \geq c^{1/\epsilon} \cdot n^{\beta}$. □