Entropy of a Zipfian Distributed Lexicon

L. C. Araujo[a], T. Cristófaro-Silva[b], H. C. Yehia[c]

[a] *Universidade Federal de São João del-Rei - Campus Alto Paraopeba - Rod. MG 443, Km 7, 36420-000, Ouro Branco, MG, Brazil*
[b] *Universidade Federal de Minas Gerais - Faculdade de Letras - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*
[c] *Universidade Federal de Minas Gerais - Departmento de Eletrônica - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*

Abstract

This article presents the calculation of the entropy of a system with Zipfian distribution. It shows that a communication system tends to present an exponent value close to, but greater than one. This choice both maximizes entropy and, at the same time, enables the retention of a feasible and growing lexicon. This result is in accordance with what is observed in natural languages and with the balance between the speaker and listener communication efforts. On the other hand, the entropy of the communicating source is very sensitive to the exponent value as well as the length of the observable data. Slight deviations on these parameters might lead to very different entropy measurements. A comparison of the estimation proposed with the entropy measure of written texts yields errors in the order of 0.3 bits and 0.05 bits for non-smoothed and smoothed distributions, respectively.

Keywords: entropy of words, Zipf's law, language communication

1. Introduction

Statistical linguistics makes use of Zipf analysis, which is a statistical tool also used in several research fields, such as economics (Mandelbrot, 1963), gene expression (Furusawa and Kaneko, 2003), and chaotic dynamic systems (Nicolis et al., 1989). Zipf found a power-law relation for the rank frequency

Email addresses: `leolca@ufsj.edu.br` (L. C. Araujo), `thaiscristofarosilva@ufmg.br` (T. Cristófaro-Silva), `hani@ufmg.br` (H. C. Yehia)

distribution of words in written texts[1] in natural languages (Zipf, 1949)

$$f_k \propto k^{-s}, \tag{1}$$

where k is the rank of a word, f_k is the word token frequency with rank k and s is the exponent of the characterizing distribution. This empirical observation has become the most remarkable statement in quantitative linguistics. The observation of Zipf-like behavior is necessary in a natural text, as much as it is the necessary behavior of any source producing information content. This is because any randomly generated symbolic sequence will follow Zipf's law with an exponent between 1 and 2 (Miller, 1957; Li, 1992). The systematic organization of language reflects the frequency usage of types. Studies have suggested that the frequency of usage is a key factor in the access of lexical items (Balota and Chumbley, 1984) and is also a driving force in language change (Bybee, 2002). Frequency plays an important role in understanding how human communication works. A useful example of Zipf analysis is given by Havlin (1995), who suggested a dissimilarity measure of two Zipf plots, from two different sources, which are smaller when the data comes from the same source and larger when it comes from different sources. This approach has been used to perform authorship attribution (Havlin, 1995).

Shannon (1948) proposed a mathematical way to deal with general communication systems and information transmission. The basic model considered by Shannon (1948) consists of i) an information source, which produces a message; ii) a transmitter, which operates on the message creating a signal suitable for transmission; iii) the channel, a mere medium where the signal is transmitted; iv) a receiver, which performs the inverse operation of the transmitter; and v) the destination, the person (or thing) for whom the message is intended. Communication is regarded as a sequence of random variables which are distributed according to the characteristics of the source. Entropy was defined as a measure of uncertainty in a random variable, what defines the expected value of the information content in a message. It is a measure of the average information produced by a source for the symbols produced in its output. Expressing entropy in bits gives us the average number of bits necessary to express each symbol produced by the source. Shannon (1948) also defined redundancy, what adds little, if any, information to a message,

[1] Text is a term used in linguistics to refer to any written or spoken passage of whatever length, that does form a unified whole (Halliday and Hasan, 1976).

but helps overcome errors arriving on the information transmission process. On a language, we might regard redundancy as a measure of the restrictions imposed on that language due to its statistical structure, which might be, for example, an expression of physiological, perceptual and phonological constraints.

The entropy of English printed words was estimated by Shannon (1951) and Grignetti (1964) using a Zipfian distribution with a characteristic exponent $s = 1$. It is known that natural languages typically present $s \approx 1$ (Piotrovskii et al., 1994). Some types of human communication still present a greater exponent, for example, children's speech has been reported to present $s \approx 1.66$ and military combat text $s \approx 1.42$. Studies on animal communication also present Zipfian behaviour. For example, McCowan et al. (1999) present an exponent value of $s \approx 1.1$ and $s \approx 0.87$ for adult and infant dolphins, respectively. The value of the exponent s seems to be related to the possible existence of a wider lexicon. Larger values of s characterize systems still in formation whereas small values characterize a well-grounded system. In this paper, we present the estimation of the entropy of a system using an arbitrary Zipfian distribution and verify the effect of the characteristic exponent s on the entropy of the system. Some results are presented to compare the estimated entropy and the entropy found in written texts, as we consider *words* as the symbols produced by a Zipfian distributed source.

2. Entropy of the System

The entropy of a system using N symbols of probabilities p_k, where $k = 1$ to N, is given by

$$\overline{H} = -\sum_{k=1}^{N} p_k \log_2 p_k = -\frac{1}{\ln 2} \sum_{k=1}^{N} p_k \ln p_k \ . \qquad (2)$$

If we consider words as the symbols used by our system, the probabilities p_k might be estimated by counting the corpus tokens and dividing it by the sample size.

George Kingsley Zipf made important contributions on language statistics, by performing word count experiments, from which he determined that there is a relationship between word frequency and rank: their product is roughly a constant (Zipf, 1949). The distribution of words in a text follows a power law:

$$p_k(s, N) = Ck^{-s} \ , \qquad (3)$$

where p_k stands for the probability of occurrence of the k-th most frequent word in the corpus; C is a normalizing constant, $C^{-1} = \sum_{n=1}^{N} n^{-s}$, which is the generalized harmonic number; k is the word rank; s is the slope, which characterizes the distribution; and N is the number of elements in the set. Zipf's law seems to hold in various languages (Zipf, 1949). "Investigations with English, Latin, Greek, Dakota, Plains Cree, Nootka (an Eskimo language), speech of children at various ages, and some schizophrenic speech have all been seen to follow this law"(Alexander et al., 1998).

Using the Zipfian value for the probabilities in Equation 2, we get to

$$\begin{aligned} \overline{H} &= -\frac{1}{\ln 2}\sum_{k=1}^{N} Ck^{-s}\ln(Ck^{-s}) \\ &= \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln k}{k^s} - \frac{\ln C}{\ln 2}\ , \end{aligned} \tag{4}$$

the summation can be calculated following the steps proposed by Grignetti (1964). We are going to find a lower and an upper bound to the entropy in Equation 4 and, to achieve that, we need to obtain bounds such that

$$B_l \leq \sum_{k=1}^{N} k^{-s}\ln k \leq B_u. \tag{5}$$

Figure 1 presents the function

$$f(x) = x^{-s}\ln x \tag{6}$$

for different values of s greater than one, which are usually found in human languages. From the first derivative of f,

$$f'(x) = x^{-s-1}(1 - s\ln x)\ , \tag{7}$$

we conclude that f is a decreasing function for $x > e^{1/s}$, which can be verified in Figure 1.

Particularly, for $s \geq 1$, the function f will be a decreasing function for $x > 3$ (for a general value of s, consider what is proposed in the next section, using $q = 0$). We might then approximate the summation using the Riemann sum approximation of an integral. The left Riemann sum S_l is an overestimate and the right Riemann sum S_r is an underestimate,

$$S_r \leq \int_a^b f(x)dx \leq S_l\ . \tag{8}$$

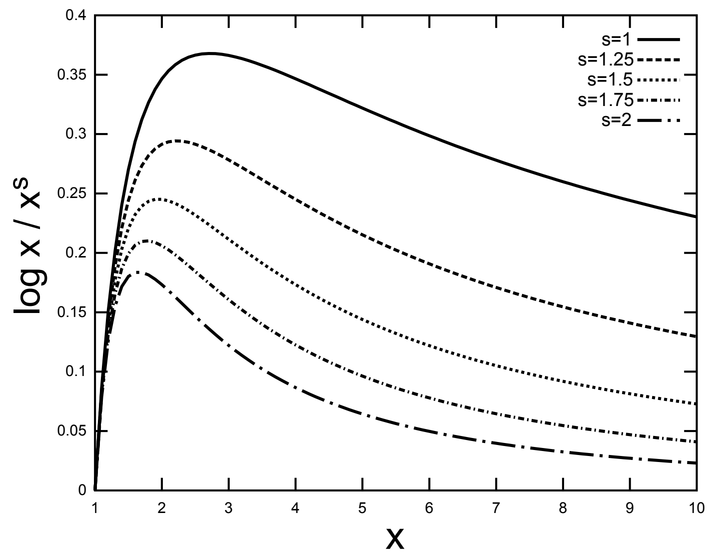

Figure 1: Function $f(x) = \ln x / x^s$ for different values of s.

Using Equation 8 we have

$$\sum_{n=4}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-2} \frac{\ln n}{n^s} \tag{9}$$

and

$$\sum_{n=4}^{N} \frac{\ln n}{n^s} \leq \int_3^{N} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \ . \tag{10}$$

From equations 9 and 10 we conclude that

$$\int_3^{N} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} \tag{11}$$

and, by adding the remaining terms (i.e. $n = 1$ and 2) to the summation, we get

$$\int_3^{N} \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} \leq \sum_{n=1}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} \ . \tag{12}$$

The wanted bounds are given by adding the N-th term to the above equation,

$$B_l = \int_3^N \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} \leq \sum_{n=1}^{N} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} = B_u \ . \tag{13}$$

Finally, the entropy bounds (H_l for the lower bound and H_u for the upper bound) are given using equations 4 and 13

$$\frac{sC}{\ln 2} B_l - \frac{\ln C}{\ln 2} = H_l \leq \overline{H} \leq H_u = \frac{sC}{\ln 2} B_u - \frac{\ln C}{\ln 2} \ . \tag{14}$$

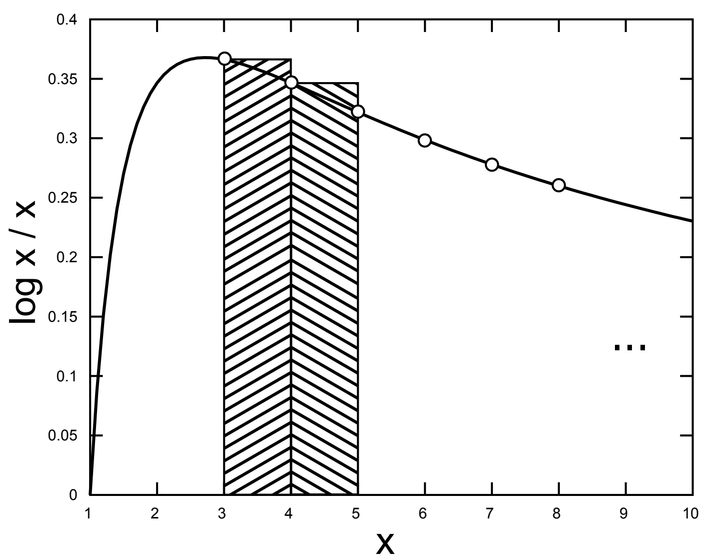

Figure 2: Left Riemann sum approximation of the integral.

The proposed approximation procedure is illustrated in Figure 2. The integral in Equation 13 is solved by parts, giving the following result, when $s \neq 1$:

$$\int \frac{\ln x}{x^s} dx = \frac{x^{1-s}}{1-s} \left(\ln x - \frac{1}{1-s} \right) \ , \tag{15}$$

where the integration constant is omitted, since it is irrelevant when evaluating the integral in an interval. When $s = 1$ the integral will result in

$$\int \frac{\ln x}{x} dx = \frac{(\ln x)^2}{2} \ . \tag{16}$$

Using Equations 4, 12 and 15 (or 16) we are able to calculate the entropy bounds of a Zipfian distributed source for a given s and N. Figure 3 presents some results for different corpora. We observe that the entropy decreases with s and increases with N.

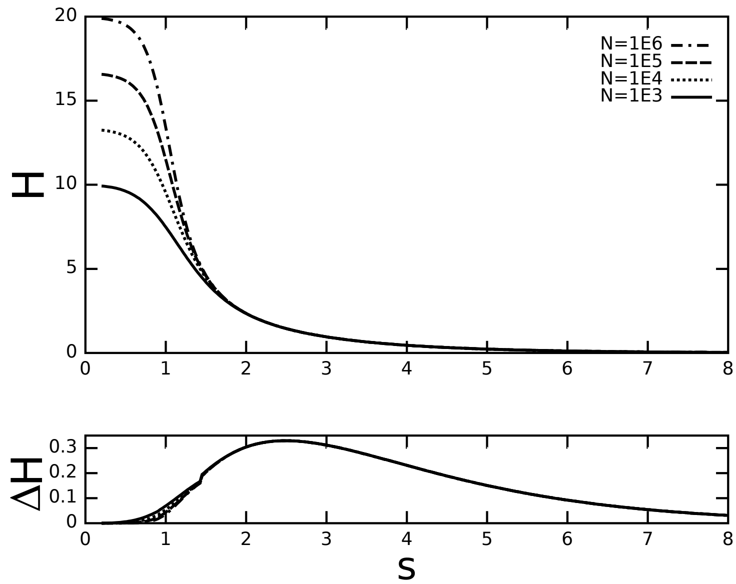

Figure 3: Entropy H (in bits) as a function of the Zipf exponent s and the number of types N. The upper plot presents the average Entropy estimated and the lower plot presents the difference between the upper and lower bounds of the entropy estimated.

The bounds for the entropy found here are rather tight (see ΔH in Figure 3), which leads to a good approximation as we consider the average of the bound values. The entropy is very sensitive to the parameter s in the vicinities of 1, where language communication is usually found. The size of the corpus also deeply influences the value of the entropy for small values of s. The proposed method states a way to estimate the entropy of theoretical Zipfian distributed sources, given its characteristic exponent value s and the number of types N.

2.1. Entropy of Real Texts

In the previous section we presented a method to estimate the entropy of a Zipfian distributed source. In this section we present the results from some measurements of entropy for the following texts: Alice in Wonderland (Lewis Carroll); Hamlet (William Shakespeare); Macbeth (William Shakespeare); The Complete Works of William Shakespeare; and Ulysses (James Joyce).

Table 1 presents the results obtained from this comparative study using the Zipf exponent calculated through a maximum likelihood estimation and the number of types found in each text. Although the estimated entropy is close to the observed entropy, it is consistently smaller, for the reasons we argue below. The estimated Zipf's exponents are also biased, since there is a flattening in low rank.

Zipf's law is a model that coarsely approximates the distribution of words in language. Two important deviations of the model from real data are observed. On the low rank region, real data usually present a flattened pattern, which was accounted for by Mandelbrot (1965). This type of flattening results in a higher entropy, since it shapes a part of the distribution towards a uniform distribution. On the long tail (high rank), a staircase pattern is usually found, due to the undersampling of rare words. This second deviation from the ideal Zipfian model does not produce a significant change on the entropy measure, therefore only the flattened pattern deviation will be considered here.

2.2. Zipf-Mandelbrot Entropy

In order to take into account the flattening observed on the low rank region of a Zipf plot, Mandelbrot (1965) introduced a modification of Zipf's law, adding a constant q to the rank k, resulting in Zipf-Mandelbrot's law

$$p_k(s, q, N) = C(k+q)^{-s} \ , \tag{17}$$

where the new normalizing constant (a generalization of a harmonic number) is given by $C^{-1} = \sum_{n=1}^{N} (n+q)^{-s}$.

Applying the same steps to this generalized formulation, the entropy will be given by

$$\overline{H} = \frac{sC}{\ln 2} \sum_{k=1}^{N} \frac{\ln(k+q)}{(k+q)^s} - \frac{\ln C}{\ln 2} \ . \tag{18}$$

The new function that will be used by the Riemmann integral approximation is

$$f(x) = (x+q)^{-s} \ln(x+q) \ , \tag{19}$$

which is decreasing for $x > e^{1/s} - q$. The constant q is a real value in the interval $[0;\infty)$. We shall then define an integer constant

$$K = \max(\lceil e^{1/s} - q \rceil, 1) \ , \tag{20}$$

which guarantees that the function $f(x)$ is decreasing for $x > K \geq 1$.

Using the left and right Riemann sum again, we find the inequalities below, which are respectively equivalent to equations 9 and 10:

$$\sum_{n=K+1}^{N-1} \frac{\ln(n+q)}{(n+q)^s} \leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s} dx \leq \sum_{n=K}^{N-2} \frac{\ln(n+q)}{(n+q)^s} \ , \tag{21}$$

$$\sum_{n=K+1}^{N} \frac{\ln(n+q)}{(n+q)^s} \leq \int_K^{N} \frac{\ln(x+q)}{(x+q)^s} dx \leq \sum_{n=K}^{N-1} \frac{\ln(n+q)}{(n+q)^s} \ . \tag{22}$$

From the above equations we conclude that

$$\int_K^{N} \frac{\ln(x+q)}{(x+q)^s} dx \leq \sum_{n=K}^{N-1} \frac{\ln(n+q)}{(n+q)^s} \leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s} dx + \frac{\ln(K+q)}{(K+q)^s} \ , \tag{23}$$

which is equivalent to Equation 11. By adding the remaining terms we get the following boundaries

$$\begin{aligned} B_l &= \int_K^{N} \frac{\ln(x+q)}{(x+q)^s} dx + \sum_{n=1}^{K-1} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} \\ &\leq \sum_{n=1}^{N} \frac{\ln(n+q)}{(n+q)^s} \\ &\leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s} dx + \sum_{n=1}^{K} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} = B_u \ . \end{aligned} \tag{24}$$

The integral in Equation 24 is solved by parts, giving the same results presented in equations 15 and 16, considering that we have $x + q$ instead of x. By adding the parameter q, the distribution suffers a flattening in the lower rank values and, consequently, the entropy of the source increases.

Table 1 also presents a comparison between the entropy estimates and the entropy found in real text data. We might observe that the usage of the Zipf-Mandelbrot model has improved the estimation of entropy. The improvement is more evident when smoothing is applied prior to the computation of the entropy.

2.3. Simple Good-Turing Smoothing

Many linguistic phenomena might be essentially regarded as infinite: words and sentences, for example. No matter how large a sample size is, it is always prudent to consider that many types have not appeared in that sample, thus they should not receive zero probability, since it is a matter of chance that they do not appear in the sample while others appear just once. Smoothing techniques reallocate the mass probability of types in order to provide a way of taking into account the probability of those types not observed in corpora. It adjusts the maximum likelihood estimates of probabilities in order to achieve a better estimate when there is insufficient data to accurately approximate them. The name *smoothing* is used because it tends to flatten the probabilities by lowering high probabilities and increasing the low ones (Chen and Goodman, 1998).

A particular technique that we use here is Good-Turing smoothing. It considers that unseen events together have a probability equal to the sum of the probabilities of all events that were observed only once, since they are equally rare. It is important to note that there is a relationship between Turing's smoothing formula and Zipf's law: both are shown to be instances of a common class of re-estimation formula and Turing's formula "smooths the frequency estimates towards a geometric distribution. (...) Although the two equations are similar, Turing's formula shifts the frequency mass towards more frequent types[2]" (Samuelsson, 1996).

The Good-Turing method states an estimation f^* for the frequency of occurrence f based on the type count for a given frequency N_f and that given frequency plus one N_{f+1}:

$$f^* = (f+1)\frac{E[N_{f+1}]}{E[N_f]} \ , \tag{25}$$

where $E[\cdot]$ represents the expectation of a random variable. The estimation f^* is usually called the "adjusted number of observations", that represents how many words you are expected to see with a given frequency of occurrence. The probability of the unseen events will be approximated by $E[N_1]/N$. The value of N_1 is the largest value and the best estimate among all other N_f. For that reason, the value of N_1 is a good approximation of the value of $E[N_1]$.

[2]We have replaced *species* in the original text by *types* which is more appropriate in the context.

One particular problem with the Good-Turing method is that, for a given f, N_{f+1} might not exist. Simple Good-Turing (SGT) (Gale, 1994) solves this problem by choosing $E[\cdot]$ so that

$$E[N_{f+1}] = E[N_f]\left(\frac{f}{f+1}\right)\left(1 - \frac{E[N_1]}{N}\right) , \tag{26}$$

leading to

$$p_f^* = p_f\left(1 - \frac{E[N_1]}{N}\right) , \tag{27}$$

as the estimated probability for types with a given frequency f. This method was shown to be accurate in a Monte Carlo study using a predefined known model and by comparing the results with other smoothing techniques (Gale, 1994).

The results in Table 1 show the difference in the entropy measures for a given corpus with and without SGT smoothing. The estimated entropy, by the method proposed here, is significantly closer to the measured entropy when SGT smoothing is applied.

Table 1: Entropy of real texts (bits), with and without SGT smoothing, compared with the estimated entropy (bits) using the parameter N (number of types) found in the text, parameter s (Zipf exponent) found by a Maximum Likelihood Estimation (MLE) and the flattening parameter q, also found by MLE.

source	N	estimated parameters			entropy		estimated entropy	
		Zipf	Zipf-Mandelbrot		normal	sgt	Zipf	Zipf-Mandelbrot
		s	s	q				
Alice	3016	0.992	1.172	3.27	8.49	8.79	8.55	8.73
Hamlet	5447	0.991	1.087	1.64	9.04	9.08	9.09	9.13
Macbeth	4017	0.969	1.009	0.56	9.00	9.00	9.02	9.04
Shakespeare	29847	1.060	1.172	2.33	9.52	9.57	9.60	9.69
Ulysses	34391	1.025	1.085	1.18	10.19	10.25	10.22	10.25

3. Conclusion

The entropy of a system with Zipfian distributed symbols decreases with the characteristic exponent s. A value of s greater than one is a necessary condition for the convergence of the generalized harmonic number. In the limit ($N \to \infty$), it is regarded as the Riemann zeta function, which converges for real $s > 1$. An exponent s which satisfies this condition leads to a Zipfian distributed lexicon which will hold regardless of how big the lexicon is.

This limiting value of s close to one is also found by Cancho and Solé (2003) when they proposed "an energy function combining the effort for the hearer and the effort for the speaker". The minimization of this function leads to a Zipfian distribution where $s = 1$, which is consistent with what is found in human languages. An exponent s greater than 1 is necessary in order to guarantee a hypothetically growing lexicon without bounds. We might then expect a greater exponent when language acquisition is still in process and a smaller exponent, closer to one, when this learning period is consolidated. Rudimentary and severely restricted communication systems might experience an exponent smaller than one, since they are not expected to evolve and widen through time, and that choice increases the entropy of the source.

The maximum rank and the repertoire size are influenced by the length of the observation but, in practical aspects, it will always be limited due to a finite observation interval. The set of words observed in the sample will always lead to a finite lexicon. An infinite lexicon is only a hypothetical approximation, which is important to analyze under the assumption of the constantly growing underlying lexicon used in human communication. Figure 4 presents an adaptation from Mandelbrot (1953), where the entropy of a finite and an infinite lexicon are compared as functions of the Zipf exponent. From both figures 3 and 4, we might observe that the length of the sample is crucial in determining the entropy of the source. A simple truncation of the sample may lead to a severe distortion of the entropy estimate. It is also important to note that the entropy estimate is much more sensitive for s in the vicinity of 1, meaning that two sources with different characteristics might have similar values of their Zipf exponent, but present quite different entropy estimates. The proposed estimation for the entropy of a Zipfian distributed source presents consistent results with real data. As the estimated measure is very sensitive to the exponent s, a slight deviation from the true exponent might lead to a poor estimation of the information associated with the communication system.

The results in Table 1 show that there is a noticeable difference between the estimated entropy of Zipf's model and the real entropy of written texts. A better approach is given by considering the generalized Zipf-Mandelbrot law. Slight differences between the entropy measure and estimations are due to small deviations in the real data from the ideal model. If the Zipf or Zipf-Mandelbrot distributions are used straightforwardly on the estimation of the entropy of the source, attention must be paid to the possible deviation from the real entropy value, since what we have proposed here is a theoretical approximation of the entropy of a truly Zipfian distributed source.

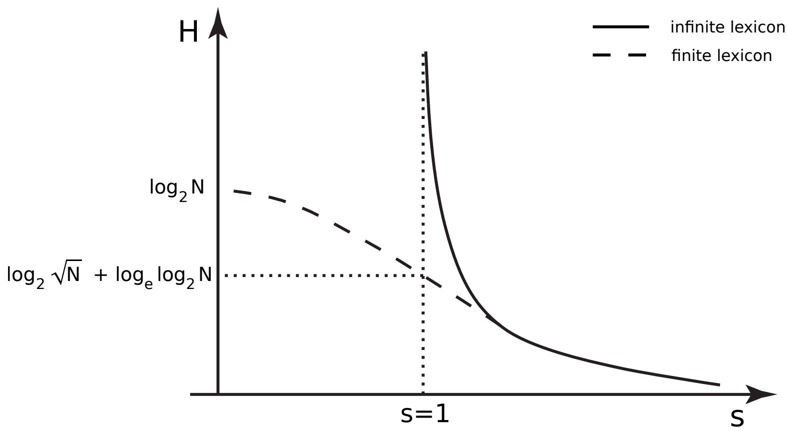

Figure 4: The entropy of a source with Zipfian distribution as a function of the characteristic exponent. Finite lexicon and infinite lexicon behaviour are compared (adapted from Mandelbrot (1953)).

Acknowledgments

This work has been supported by the Brazilian agencies CNPq and FAPEMIG.

References

Alexander, L., Johnson, R., Weiss, J., 1998. Exploring zipf's law. Teaching Mathematics Applications 17 (4), 155–158.

Balota, D. A., Chumbley, J. I., Jun. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of experimental psychology. Human perception and performance 10 (3), 340–357.

Bybee, J., 2002. Word Frequency and Context of Use in the Lexical Diffusion of Phonetically Conditioned Sound Change. Language Variation and Change 14 (3), 261–290.

Cancho, R. F., Solé, R. V., 2003. Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences of the United States of America 100 (3), 788–791.

Chen, S. F., Goodman, J., 1998. An empirical study of smoothing techniques for language modeling. Tech. rep., Computer Science Group, Harvard University.

Furusawa, C., Kaneko, K., 2003. Zipf's law in gene expression. Physical review letters 90 (8).

Gale, W., 1994. Good-Turing smoothing without tears. Journal of Quantitative Linguistics 2.

Grignetti, M., 1964. A note on the entropy of words in printed english. Information and Control 7, 304–306.

Halliday, M. A. K., Hasan, R., 1976. Cohesion in English. English language series. Longman.

Havlin, S., 1995. The distance between zipf plots. Physica A: Statistical Mechanics and its Applications 216 (1), 148–150.

Li, W., 1992. Random texts exhibit zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory 38 (6), 1842–1845.

Mandelbrot, B., 1953. Contribution à la théorie mathématique des jeux de communication. Publications de l'Institut de Statistique de l'Université de París. Institut Henri Poincaré.

Mandelbrot, B., March 1963. Oligopoly, mergers, and the paretian size distribution of firms. External research note: Nc-246, IBM.

Mandelbrot, B., 1965. Information theory and psycholinguistics. In: Wolman, B. B., Nagel, E. N. (Eds.), Scientific Psychology: Principles and Approaches. Basic Books, pp. 550–562.

McCowan, B., Hanser, S. F., Doyle, L. R., 1999. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. Animal Behaviour 57, 409–419.

Miller, G. A., Jun 1957. Some effects of intermittent silence. The American Journal of Psychology 70 (2), 311–314.

Nicolis, G., Nicolis, C., Nicolis, J. S., 1989. Chaotic dynamics, Markov partitions, and Zipf's law. Journal of Statistical Physics 54 (54), 915–924.

Piotrovskii, R. G., Pashkovskii, V. E., Piotrovskii, V. R., 1994. Psychiatric linguistics and automatic text processing. Nauchno-Tekhnicheskaya Informatsiya, Seriya 2 28 (11), 21–25.

Samuelsson, C., 1996. Relating Turing's formula and zipf's law. CoRR cmp-lg/9606013.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal.

Shannon, C. E., 1951. Prediction and entropy of printed english. Tech. Rep. 30, The Bell System Technical Journal.

Zipf, G. K., 1949. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Hafner Pub. Co.