Entropy of a Zipfian Distributed Lexicon

Leonardo Carneiro Araújo*
Campus Alto Paraopeba, Universidade Federal de São João del-Rei, Rod. MG 443, Km 7, Ouro Branco, MG, Brazil

Thaïs Cristófaro-Silva†
Faculdade de Letras, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil

Hani Camille Yehia‡
Departmento de Eletrônica, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil

(Dated: November 19, 2026)

This article presents the calculation of the entropy of a system with Zipfian distribution and shows that a communication system tends to present an exponent value close to one, but still greater than one, so that it might maximize entropy and hold a feasible lexicon with an increasing size. This result is in agreement 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, making it a poor parameter to characterize the communication process.

arXiv:1307.0898v2 [stat.AP] 18 Nov 2026

I. INTRODUCTION

Statistical linguistics makes uses of Zipf analysis, which is a statistical tool used as well in several research fields, such as economics [1], gene expression [2], and chaotic dynamic systems [3]. Zipf found a power-law relation for written texts in natural languages [4]. This empirical observation has become the most remarkable statement in quantitative linguistics. The observation of a Zipf behavior is necessary in a natural text, as much as it is a necessary behavior of any source producing information contents, since any randomly generated symbolic sequence will present a Zipf's law with an exponent between 1 and 2 [5, 6]. The systematic organization of language reflects the frequencies of usage of types. Studies have suggested that the frequency of usage is a key factor in the access of lexical items [7] and also a driving force in language change [8]. Frequency plays an important role in understanding how human communication works. A useful example of Zipf analysis is given by [9], who suggested a dissimilarity measure of two Zipf plots, from two different sources, which will be smaller when the data come from the same source and larger when they come from different sources. This approach is used to perform authorship attribution [9].

The definitions of entropy and redundancy of a language were introduced by [10]. Entropy is a measure of the average information produced by a source for each symbol produced in its output. Expressing the entropy in bits gives us the average number of bits necessary to express each produced by the source. Redundancy measures the restrictions imposed on a language due to its statistical structure, what might be an expression of physiological and phonological constraints.

The entropy of English printed words was estimated by [11] and [12] using a Zipfian distribution with a characteristic exponent $s = 1$. It is known that natural languages typically present $s \approx 1$ [13]. Some types of human communications still present a greater exponent, for example, child speech has been reported to present $s \approx 1.66$ and military combat text $s \approx 1.42$. Studies on animal communication also present a Zipf's behaviour, for example, [14] 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 describe the plasticity of the communication system, what leads to a potentially growing lexicon. Larger values of s characterize systems still in formation and small values systems well-grounded. In this paper, we are going to present the calculation 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.

II. ENTROPY OF THE SYSTEM

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

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

If we consider words as the symbols used by our system, the probabilities p_k might be estimated by counting the frequency of occurrence of types and dividing it by our sample size. No matter how large our corpus is, there

* leolca@ufsj.edu.br
† thaiscristofarosilva@ufmg.br
‡ hani@ufmg.br

might always be words in the underlying lexicon that have no representation on that corpus. In order to acquire a better estimate of the probabilities, we should perform a Turing Smoothing [15, 16].

George Kingsley Zipf made important contributions on language statistics, performing word count experiments, from which he determined that there is a relationship between word's frequency of appearance in texts and its rank, the product of them being roughly a constant [4]. The distribution of words in a language follows a power law: $p_k(s,N) = Ck^{-s}$, where p_k stands for 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 the slope, which characterizes the distribution; and N is the number of elements in the set.

Zipf's law seems to hold regardless the language observed [4]. "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"[17]. Since a smoothing is necessary, in order to achieve a better approximation of the underlying probabilities, it is important to notice that there is a relation between the 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 species"[18].

Using the Zipfian value for the probabilities in Equation 1, we get

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

where the summation might be calculated following the steps proposed by [12]. Figure 1 presents the function $f(x) = x^{-s}\ln x$ for different values of s greater than one, which is usually found in human languages. From the first derivative of f,

$$f'(x) = x^{-s-1}(1 - s\ln x)\ , \tag{3}$$

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

Particularly for $s \geq 1$, the function f will be a decreasing function for $x > 3$. 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,

$$S_r \leq \int_a^b f(x)dx \leq S_l\ . \tag{4}$$

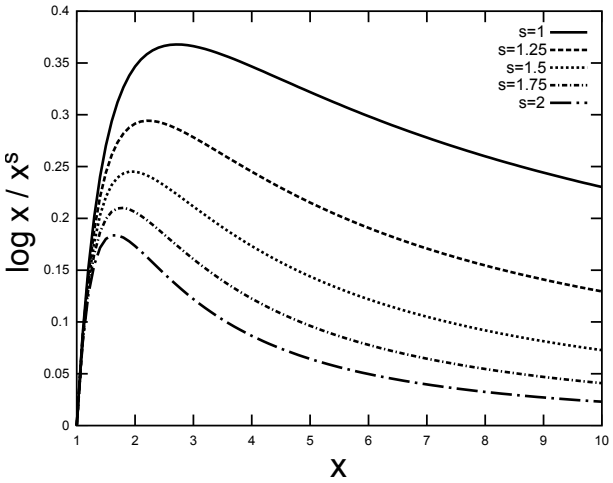


FIG. 1. Function $f(x) = \ln x/x^s$ for different values of s.

Using Equation 4 we might write

$$\sum_{n=4}^{N-1}\frac{\ln n}{n^s} \leq \int_3^{N-1}\frac{\ln x}{x^s}dx \leq \sum_{n=3}^{N-2}\frac{\ln n}{n^s} \tag{5}$$

and

$$\sum_{n=4}^{N}\frac{\ln n}{n^s} \leq \int_3^{N}\frac{\ln x}{x^s}dx \leq \sum_{n=3}^{N-1}\frac{\ln n}{n^s}\ . \tag{6}$$

From Equations 5 and 6 we conclude that

$$\int_3^N \frac{\ln x}{x^s}dx \leq \sum_{n=3}^{N-1}\frac{\ln n}{n^s} \leq \int_3^{N-1}\frac{\ln x}{x^s}dx + \frac{\ln 3}{3^s} \tag{7}$$

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

$$\int_3^N \frac{\ln x}{x^s}dx + \frac{\ln 2}{2^s} \leq \sum_{n=1}^{N-1}\frac{\ln n}{n^s} \leq \int_3^{N-1}\frac{\ln x}{x^s}dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s}\ . \tag{8}$$

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

$$\begin{aligned} \int_3^N \frac{\ln x}{x^s}dx + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} &\leq \sum_{n=1}^{N}\frac{\ln n}{n^s} \\ &\leq \int_3^{N-1}\frac{\ln x}{x^s}dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s}\ . \end{aligned} \tag{9}$$

The proposed approximation procedure is illustrated in Figure 2. The integral in Equation 9 is solved by parts, giving the following result, when $s \neq 1$:

$$\int \frac{\ln x}{x^s}dx = \frac{x^{1-s}}{1-s}\left(\ln x - \frac{1}{1-s}\right)\ , \tag{10}$$

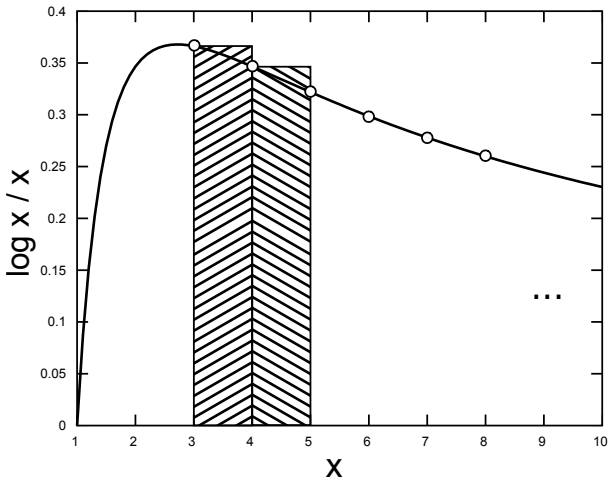


FIG. 2. Left Riemann sum approximation of the integral.

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

$$\int \frac{\ln x}{x} dx = \frac{(\ln x)^2}{2} \ . \tag{11}$$

Using Equations 2, 9 and 10 (or 11) we are able to calculate the bounds of the entropy of a Zipfian distributed source for a given s and N. Figure 3 presents some results for different corpora lengths. We might observe that the entropy decreases with s and increases with N.

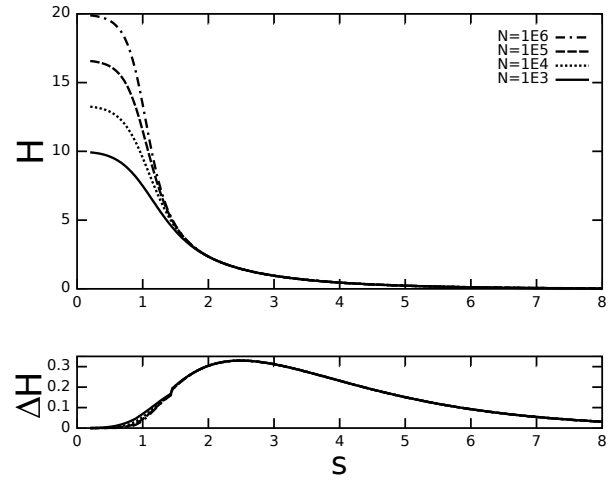


FIG. 3. Entropy H (in bits) as a function of the Zipf exponent s and sample length N. The upper plot presents the average Entropy estimated and the lower plot presents the difference between the upper and lower bounds of the entropy estimated.

III. CONCLUSION

The entropy of a system with Zipfian distributed symbols is decreasing 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, 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 distribution of the lexicon which will hold regardless how big the lexicon is.

This limiting value of s close to one is also found by [19] 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 with $s = 1$, which is consistent with what is found in human languages. An exponent s greater than 1 is necessary to guarantee an hypothetical 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 consolidating. Rudimentary and severely restricted communication processes might experience an exponent smaller than one, since they are not expected to evolve and widen through time, and that choice maximizes the entropy of the source.

The maximum rank and the repertoire size are influenced by the length of our observation but, in practical aspects, it will always be limited due to our finite observation interval. It will always lead to a finite lexicon, the set of words observed in the sample. 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 communications. Figure 4 presents an adaptation from [20], where the entropy of an finite and infinite lexicon are compared as a function of the Zipf exponent. From both figures 3 and 4, we might observe that the length of the observation sample is crucial in determining the entropy of the source. A simple truncation on the sample may lead to a severe distortion on 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 exponent s might then be a poor parameter to characterize the information in a communication process whose symbols are Zipfian distributed. Therefore, if the Zipf distribution is used straightforwardly in the estimation of the entropy of the source, attention must be paid to its actual meaning

ACKNOWLEDGMENTS

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

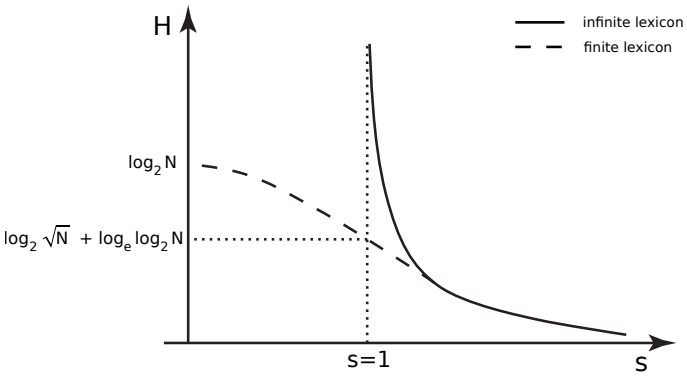


FIG. 4. The entropy of a source with Zipfian distribution as a function of the characteristic exponent. A finite lexicon and an infinite lexicon behaviours are compared (adapted from [20]).

[1] B. Mandelbrot, *Oligopoly, Mergers, and the Paretian Size Distribution of Firms*, External Research Note: NC-246 (IBM, 1963).

[2] C. Furusawa and K. Kaneko, Physical review letters **90** (2003).

[3] G. Nicolis, C. Nicolis, and J. S. Nicolis, Journal of Statistical Physics **54**, 915 (1989).

[4] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Hafner Pub. Co, 1949).

[5] G. A. Miller, The American Journal of Psychology **70**, 311 (1957).

[6] W. Li, IEEE Transactions on Information Theory **38**, 1842 (1992).

[7] D. A. Balota and J. I. Chumbley, Journal of experimental psychology. Human perception and performance **10**, 340 (1984).

[8] J. Bybee, Language Variation and Change **14**, 261 (2002).

[9] S. Havlin, Physica A: Statistical Mechanics and its Applications **216**, 148 (1995).

[10] C. E. Shannon, Bell System Technical Journal (1948).

[11] C. E. Shannon, *Prediction and entropy of printed English*, Tech. Rep. 30 (The Bell System Technical Journal, 1951).

[12] M. Grignetti, Information and Control **7**, 304 (1964).

[13] R. G. Piotrovskii, V. E. Pashkovskii, and V. R. Piotrovskii, Nauchno-Tekhnicheskaya Informatsiya, Seriya 2 **28**, 21 (1994).

[14] B. McCowan, S. F. Hanser, and L. R. Doyle, Animal Behaviour **57**, 409 (1999).

[15] I. J. Good, Biometrika **40(3-4)**, 237 (1953).

[16] W. Gale, Journal of Quantitative Linguistics **2** (1994).

[17] L. Alexander, R. Johnson, and J. Weiss, Teaching Mathematics Applications **17**, 155 (1998).

[18] C. Samuelsson, CoRR **cmp-lg/9606013** (1996).

[19] R. F. Cancho and R. V. Solé, Proceedings of the National Academy of Sciences of the United States of America **100**, 788 (2003).

[20] B. Mandelbrot, *Contribution à la théorie mathématique des jeux de communication*, Publications de l'Institut de Statistique de l'Université de París (Institut Henri Poincaré, 1953).

[21] B. Mccowan, L. R. Doyle, J. M. Jenkins, and S. F. Hanser, Animal Behaviour **69**, F1 (2005).

[22] B. Mandelbrot, in *Scientific Psychology: Principles and Approaches*, edited by B. B. Wolman and E. N. Nagel (Basic Books, 1965) pp. 550–562.

[23] S. F. Chen and J. Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling*, Tech. Rep. (Computer Science Group, Harvard University, 1998).

[24] M. A. K. Halliday and R. Hasan, *Cohesion in English*, English language series (Longman, 1976).