Advances in Complex Systems

ZIPF'S LAW AND RANDOM TEXTS

RAMON FERRER I CANCHO*
Complex Systems Research Group, FEN
Universitat Politècnica de Catalunya
Campus Nord B4, 08034 Barcelona, Spain
E-mail: ramon@complex.upc.es

RICARD V. SOLÉ
(1) Complex Systems Research Group, FEN
Universitat Politècnica de Catalunya
Campus Nord B4, 08034 Barcelona, Spain
E-mail: ricard@complex.upc.es

(2) Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA

Received (received date)
Revised (revised date)

Random-text models have been proposed as an explanation for the power law relationship between word frequency and rank, the so-called Zipf's law. They generally regarded as null hypothesis rather than models in the strict sense. In this context, recent theories of language emergence and evolution assume this law as *a priori* information with no need of explanation. Here random texts and real texts are compared through (a) so-called lexical spectrum and (b) the distribution of words having the same length. It is shown that real texts fill the lexical spectrum much more efficiently and regardless of the word length, suggesting that Zipf's law meaningfulness is high.

Understanding the origins and evolution of language requires an appropriate identification of its universal features. One of the most obvious is the statistical distribution of word abundances. Word frequency distributions exhibit striking regularities. If words in a sample text are ordered decreasingly by their frequency, the (normalized) frequency of a word is a power law of its rank [16], r, described in its simplest form as

$$P(r) \propto r^{-\alpha} \tag{0.1}$$

where $P(r)$ is the normalized frequency of a word whose rank is r. Equivalently, such regularity can be presented (again in its simplest form) as a function of the

*corresponding author.

frequency f of a word and becomes

$$P(f) \propto f^{-\beta} \tag{0.2}$$

where $P(f)$ is the probability a word has frequency f in a sample. The second form is called the lexical spectrum [15] or the inverse Zipf's distribution [3]. The exponents in Eq. 0.1 and 0.2 obey (see for instance [9,10,6]

$$\beta = \frac{1}{\alpha} + 1 \tag{0.3}$$

and their typical values are $\alpha = 1$ and $\beta = 2$. Both Eq. 0.1 and 0.2 are the so-called Zipf's law, although the former is the most common one. Here, the term law refers to the strength of the empirical observation, that has been tested in different languages and authors [1]. As far as we know, detailed and extensive study has only shown that the values of the exponents can vary from one sample to another [1] and even more than one domain [15,12,6] is necessary for explaining the same sample. Fig. 1 shows the normalized frequency versus rank ($\alpha = 1$) and the lexical spectrum ($\beta = 2$) for Herman Melville's Moby Dick.

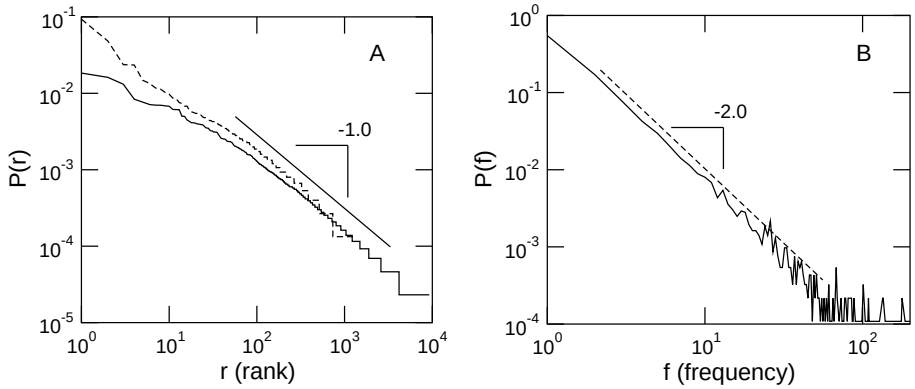

Fig. 1. Frequency versus rank (A) and lexical spectrum (B) of Herman Melville's Moby Dick (9,244 different words). The dashed line in A shows the frequency versus rank for words having length 5, which is the average length of words in Melville's book (there are 1,248 different 5-letter words). The exponents are (A) $\alpha = 1$ and (B) $\beta = 2 = \frac{1}{\alpha} + 1$ as expected.

One obvious question raised by these observations is: are they the result of some non trivial causal process? Any observed regularity in nature needs first to be studied by means of null models. One possible explanation of the Zipf's law comes from a purely random process. An early argument against any special causal explanation beyond randomness was the discovery that random sequences of letters (in which the blank space was among them) reproduced the $\alpha = 1$ exponent of words [7]. Assume that the keys of a typewriter are typed at random. If the blank space is hit with probability q and one of the N possible letters are hit with probability

$(1-q)/N$, having all letters the same probability, the distribution of words limited by blank spaces can be shown to obey Eq. 0.1 [7]. Ref. [5] provides a recent proof when all characters have the same probability. Such a random process is called a *monkey language* [7,2] or a *intermittent silence* [7] or simply a *random text* [5] model. Despite of the surprising suitability of random texts, they are generally regarded as null hypothesis that whatever explanation has to face [8]. To some extent, it has been concluded and widespread that Zipf's law does not tell anything (relevant) about language [13,5].

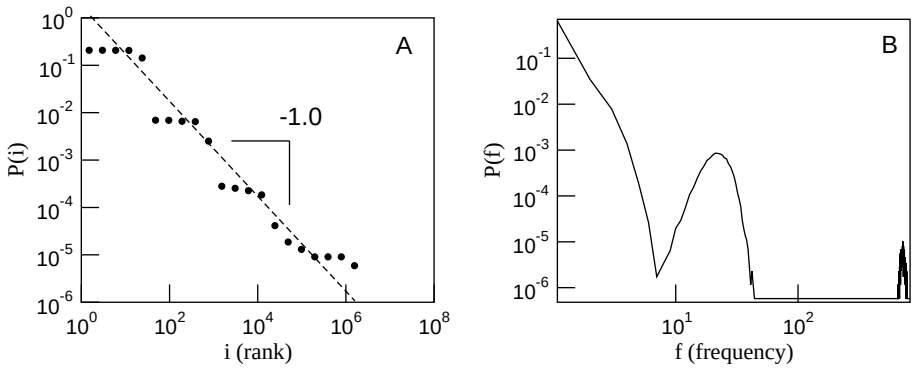

Fig. 2. Frequency versus rank (A) and lexical spectrum (B) of a so-called monkey language (a random text) formed by $1,731,411$ different words ($4 \cdot 10^6$ total words). The alphabet has $N = 26$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A) is $\alpha = 1$ while no power law seems to fit in (B).

Such a conclusion comes, in our view, from the misleading comparison between rank distributions. When the lexical spectrum is plotted for the monkey language, the differences between random and non random sequences become dramatic. Fig. 2 shows the normalized frequency versus rank and the lexical spectrum for a monkey language with $q = 0.18$ and $n = 26$. The former shows $\alpha = 1$. The later should show an exponent $\beta = 2$ as predicted by Eq. 0.3 but no power domain can be identified and it differs greatly from its counterpart in Fig. 1. It tempting to think the statistical structure of both distributions is completely different.

Vocabulary growth in random texts is faster than in real texts [3]. $N = 26$ leads to $11,881,376$ different 5-letter words, far from the about $1,7$ million words of the random text in Fig 2. If sampling effects are responsible for the surprising plot in Fig. 2 B, the lexical spectrum with $N = 2$ should improve (there are only 32 different 5-letter words). Fig. 3 shows that not only the frequency versus rank plot improves but also the lexical spectrum. Nonetheless, the quality of the latter is still clearly lower than that of a real text. The analytically predicted exponents are obviously valid in Fig. 2-3 B but random texts like these reveal high sampling sensitiveness when compared to real texts.

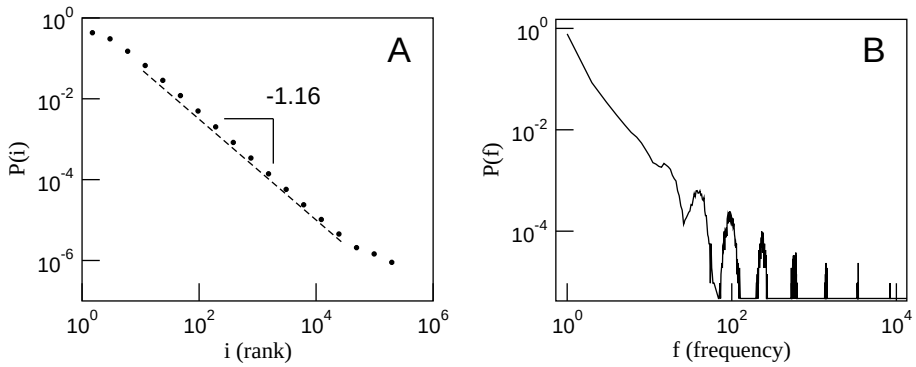

Fig. 3. Frequency versus rank (A) and lexical spectrum (B) of a so-called monkey language (a random text) formed by 212,197 different words ($4 \cdot 10^6$ total words). The alphabet has $N = 2$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A) is $\alpha = 1$ and the quality of the lexical spectrum it higher than with $N = 26$.

It might be thought the monkey language we have employed is simplistic. All letters have the same probability, which is not realistic. If a random text is generated with letter probabilities obtained from Moby Dick, the frequency versus rank plot loses the steps-like appearance (solid line in Fig. 4 A) while the lexical spectrum improves (Fig. 4 B). Notice that the improvement can not be attributed to a smaller vocabulary (about 1.7 million word in the unbiased case) but a less restrictive way of filling the spectrum.

An additional source of disagreement comes from the analysis of word distributions of a certain length. Monkey languages imply word length follows an exponential distribution given by

$$P(L) \propto (1-q)^L \tag{0.4}$$

where $P(L)$ is the probability of words formed by L letters. In contrast, word length is modeled with log-normal [1,11] or Poissonian distributions [4]. Empirical studies show that there is a typical length $L > 1$ and long tails may appear. If all letters have the same probability, monkey languages predict that words having the same length have the same frequency. The dashed line in Fig. 1 A shows the distribution of words in Melville's Moby Dick having the same length, which is clearly Zipfian. In contrast, the equivalent in a monkey language in which all letters have the same probability is a uniform distribution and the distribution of letters of a monkey language with realistic letter frequencies is the dashed line in Fig. 4 A. Both are clearly not Zipfian.

By assuming that Zipf's law is a trivial statistical regularity, some authors declined to include it as part of the features of language origins. Instead, it has been used as a given statistical fact with no need for explanation [14]. Our observations do not give support to this view.

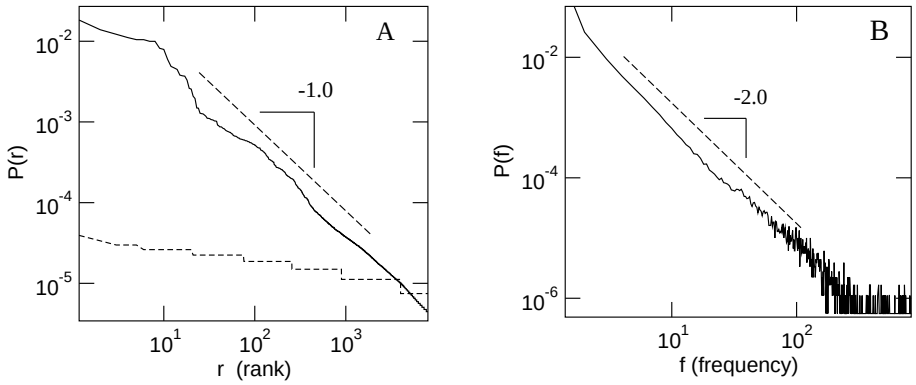

Fig. 4. Frequency versus rank (A) and lexical spectrum (B) of a monkey language formed by $1,795,617$ different words ($4 \cdot 10^6$ total words). Character probabilities were obtained from Melville's Moby Dick. The dashed line in A shows the frequency versus rank for words having length 5, which is the average length of words in Melville's book. The random text has $238,891$ different 5-letter words. The exponent in (A) is $\alpha = 1$ while the slope in (B) is $\alpha = 2.0$.

We have also shown that random texts lose the Zipfian shape in the frequency versus rank plot when words are restricted to a certain length, which is not the case of real texts. It is thus clear that monkey languages partial validity relies on their word length distribution, which we have pointed to be unrealistic. These results thus suggest that future theories of language origin should be able to explain the origin of the Zipf's law, instead of using it as a given constraint.

Acknowledgments

We thanks R. Rousseau and G. Miller for helpful comments. This work was supported by the Santa Fe Institute (RVS) and grants of the Generalitat de Catalunya (FI/2000-00393, RFC) and the CICYT (PB97-0693, RVS).

References

[1] V. K. Balasubrahmanyan and S. Naranan. Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, 3(3):177–228, 1996.

[2] J. L. Casti. Bell curves and monkey languages. *Complexity*, 1(1), 1995.

[3] A. Cohen, R. N. Mantegna, and S. Havlin. Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, 5(1):95–104, 1997.

[4] R. G. G. Wimmer, R. Köhler and G. Altmann. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1:98–106, 1994.

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

[6] R. Ferrer i Cancho and R. V. Solé. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 2001. in press.

[7] G. A. Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70:311–314, 1957.

[8] G. A. Miller and N. Chomsky. Finitary models of language users. In R. D. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 2. Wiley, New York, 1963.

[9] S. Naranan. Statistical laws in information science, language and system of natural numbers: some striking similarities. *Journal of Scientific and Industrial Research*, 51:736–755, 1992.

[10] S. Naranan and V. Balasubrahmanyan. Information theoretic models in statistical linguistics - part i: A model for word frequencies. *Current Science*, 63:261–269, 1992.

[11] S. Naranan and V. Balasubrahmanyan. Information theoretic models in statistical linguistics - part ii: Word frequencies and hierarchical structure in language. *Current Science*, 63:297–306, 1992.

[12] S. Naranan and V. Balasubrahmanyan. Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics*, 5(1-2):35–61, 1998.

[13] M. A. Nowak. The basic the reproductive ratio of a word, the maximum the size of the lexicon. *J. theor. Biol.*, 204:179–189, 2000.

[14] M. A. Nowak, J. B. Plotkin, and V. A. Jansen. The evolution of syntactic communication. *Nature*, 404:495–498, 2000.

[15] J. Tuldava. The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, 3(1):38–50, 1996.

[16] G. K. Zipf. *Human behaviour and the principle of least effort. An introduction to human ecology*. Hafner reprint, New York, 1972. 1st edition: Cambridge, MA: Addison-Wesley, 1949.