

ZIPF'S LAW AND RANDOM TEXTS

RAMON FERRER i CANCHO*[‡] and RICARD V. SOLÉ^{†,‡,§}

[‡]*Complex Systems Research Group, FEN, Universitat Politècnica de Catalunya,
Campus Nord B4, 08034 Barcelona, Spain*

and

[§]*Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA*

**ramon@complex.upc.es*

†ricard@complex.upc.es

Received 8 October 2001

Revised 22 October 2001

Accepted 23 October 2001

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 are generally regarded as null hypotheses 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) the 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 the meaningfulness of Zipf's law is high.

Keywords: Human language; scaling; Zipf's law; monkey languages; random texts.

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

$$P(r) \propto r^{-\alpha}, \quad (1)$$

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 frequency f of a word and becomes

$$P(f) \propto f^{-\beta}, \quad (2)$$

*Corresponding author.

where $P(f)$ is the probability a word has frequency f in a sample. The second form is called the lexical spectrum [14] or the inverse Zipf's distribution [3]. The exponents in Eqs. (1) and (2) obey (see for instance Refs. 4, 8 and 9)

$$\beta = \frac{1}{\alpha} + 1 \quad (3)$$

and their typical values are $\alpha = 1$ and $\beta = 2$. Both Eqs. (1) and (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 by different 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 [4, 11, 14] is necessary to explain the same sample. Figure 1 shows the normalized frequency versus rank ($\alpha = 1$) and the lexical spectrum ($\beta = 2$) for Herman Melville's *Moby Dick*.

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 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 $(1 - q)/N$, with all letters having the same probability, the distribution of words limited by blank spaces can be shown to obey Eq. (1) [7]. Reference 5 provides recent proof when all characters have the same probability. Such a random process is called a *monkey language* [2, 7] or *intermittent silence* [7] or simply a *random text* [5] model.

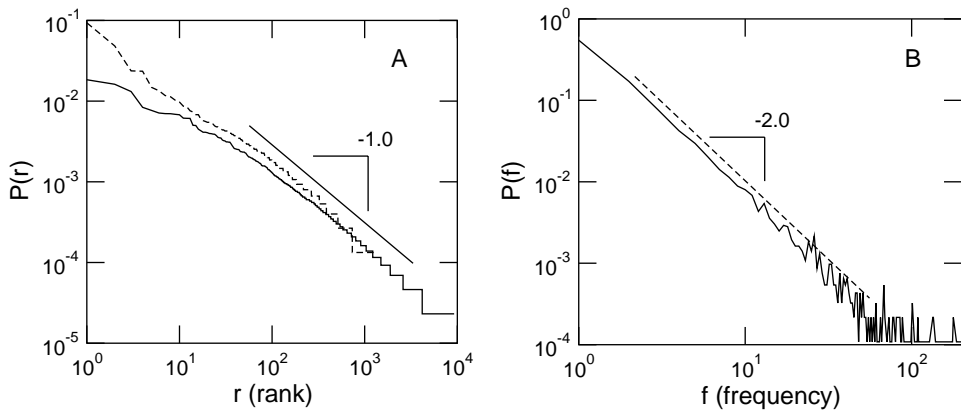


Fig. 1. Frequency versus rank (A) and lexical spectrum (B) of Herman Melville's *Moby Dick* (9,244 different words). The dashed line in (A) shows the frequency versus rank for words having length 5, which is the average length of words in Melville's book (there are 1,248 different 5-letter words). The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

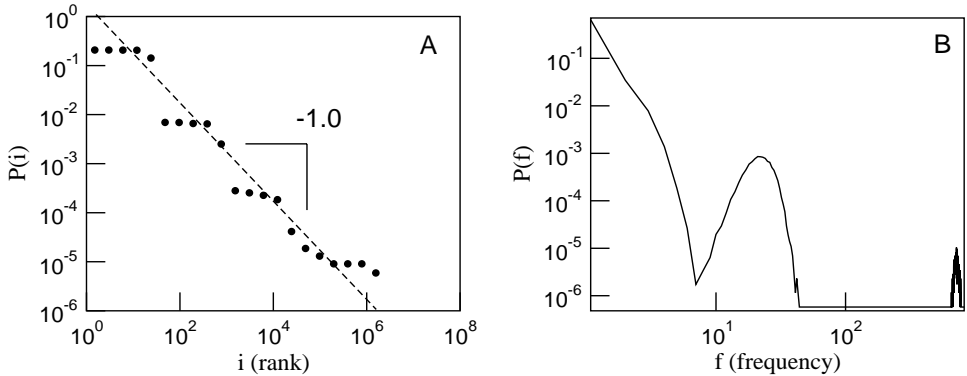


Fig. 2. Frequency versus rank (A) and lexical spectrum (B) of a so-called monkey language (a random text) formed by 1,731,411 different words (4×10^6 total words). The alphabet has $N = 26$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A)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

Since random texts easily generate statistically consistent results, they are generally regarded as a null hypothesis [6]. A consequence of this view is that Zipf's law does not provide, by itself, any meaningful information concerning language structure and its evolution [5, 13].

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. Figure 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 latter should show an exponent $\beta = 2$ as predicted by Eq. (3) but no power domain can be identified and it differs greatly from its counterpart in Fig. 1. It is tempting to think that 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 roughly 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). Figure 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 Figs. 2(B) and 3(B) but random texts like these reveal high sampling sensitivity when compared to real texts.

It might be thought that 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 its step-like appearance (solid line in Fig. 4(A)) while the lexical

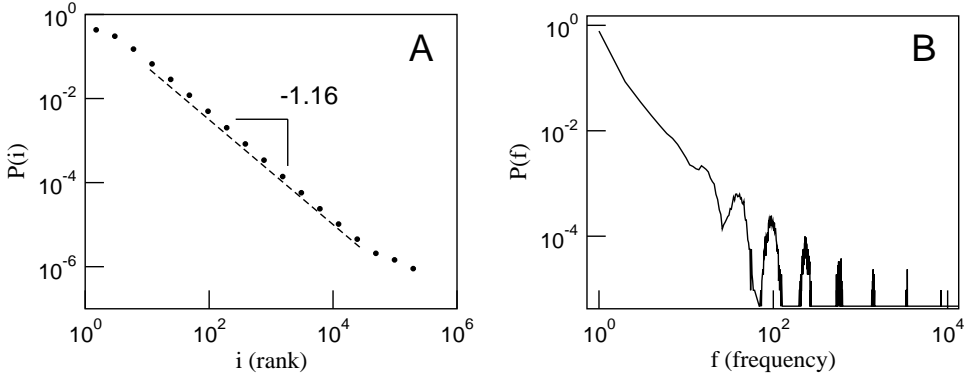


Fig. 3. Frequency versus rank (A) and lexical spectrum (B) of a so-called monkey language (a random text) formed by 212,197 different words (4×10^6 total words). The alphabet has $N = 2$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A) is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

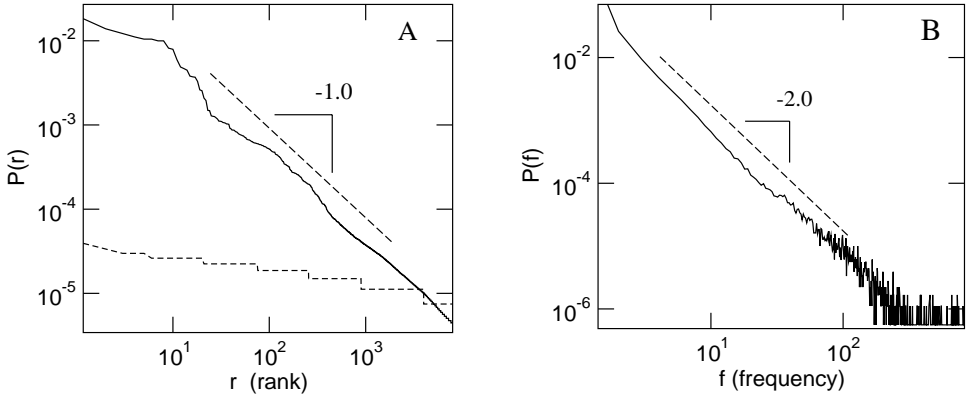


Fig. 4. Frequency versus rank (A) and lexical spectrum (B) of a monkey language formed by 1,795,617 different words (4×10^6 total words). Character probabilities were obtained from Melville's *Moby Dick*. The dashed line in (A) shows the frequency versus rank for words having length 5, which is the average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

spectrum improves (Fig. 4(B)). Notice that the improvement can not be attributed to a smaller vocabulary (about 1.7 million words 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

$$P(L) \propto (1 - q)^L, \quad (4)$$

where $P(L)$ is the probability of words formed by L letters. In contrast, word length is modeled with log-normal [1, 10] or Poissonian distributions [15]. 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 have declined to include it as part of the features of language origin. Instead, it has been used as a given statistical fact with no need for explanation [12]. Our observations do not give support to this view.

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 in real texts. It is thus clear that monkey languages' partial validity relies on their word length distribution, which we have indicated is unrealistic. These results suggest that future theories of language origin should be able to explain the origin of Zipf's law, instead of using it as a given constraint.

Acknowledgments

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

References

- [1] Balasubrahmanyam, V. K. and Naranan, S., Quantitative linguistics and complex system studies, *J. Quant. Linguistics* **3** (3), 177–228 (1996).
- [2] Casti, J. L., Bell curves and monkey languages *Complexity* **1** (1) (1995).
- [3] Cohen, A., Mantegna, R. N. and Havlin, S., Numerical analysis of word frequencies in artificial and natural language texts, *Fractals* **5** (1), 95–104 (1997).
- [4] Ferrer i Cancho, R. and Solé, R. V., Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited, *J. Quant. Linguistics* (2001), in press.
- [5] Li, W., Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Trans. Inf. Theory* **38** (6), 1842–1845 (1992).
- [6] Miller, G. A. and Chomsky, N., Finitary models of language users, in *Handbook of Mathematical Psychology*, Vol. 2, Luce, R. D., Bush, R. and Galanter, E., eds. (Wiley, New York, 1963).
- [7] Miller, G. A., Some effects of intermittent silence, *Amer. J. Psychology* **70**, 311–314 (1957).
- [8] Naranan, S., Statistical laws in information science, language and system of natural numbers: Some striking similarities, *J. Scientific and Industrial Research* **51**, 736–755 (1992).
- [9] Naranan, S. and Balasubrahmanyam, V., Information theoretic models in statistical linguistics — Part I: A model for word frequencies, *Current Science* **63**, 261–269 (1992).

- [10] Naranan, S. and Balasubrahmanyam, V., Information theoretic models in statistical linguistics — Part II: Word frequencies and hierarchical structure in language, *Current Science* **63**, 297–306 (1992).
- [11] Naranan, S. and Balasubrahmanyam, V., Models for power law relations in linguistics and information science, *J. Quant. Linguistics* **5** (1&2), 35–61 (1998).
- [12] Nowak, M. A., Plotkin, J. B. and Jansen, V. A., The evolution of syntactic communication, *Nature* **404**, 495–498 (2000).
- [13] Nowak, M. A., The basic the reproductive ratio of a word, the maximum the size of the lexicon. *J. Theor. Biol.* **204**, 179–189 (2000).
- [14] Tuldava, J., The frequency spectrum of text and vocabulary, *J. Quant. Linguistics* **3** (1), 38–50 (1996).
- [15] Wimmer, R. G. G., Köhler, R. and Altmann, G., Towards a theory of word length distribution, *J. Quant. Linguistics* **1**, 98–106 (1994).
- [16] Zipf, G. K., *Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology*, Hafner reprint, New York, 1972, 1st edn. (Addison-Wesley, Cambridge, MA, 1949).