Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited

Ramon Ferrer Cancho*[1] and Ricard V. Solé*,+

* Complex Systems Research Group, FEN, UPC
Campus Nord, B4-B5, Barcelona 08034 SPAIN

+Santa Fe Institute, 1399 Hyde Park Road,
New Mexico 87501, USA

Abstract

Zipf's law states that the frequency of a word is a power function of its rank. The exponent of the power is usually accepted to be close to $(-)1$. Great deviations between the predicted and real number of differents words of a text, disagreements between the predicted and real exponent of the probability density function and statistics on a big corpus, make evident that word frequency as a function of the rank follows two different exponents, $\approx (-)1$ for the first regime and $\approx (-)2$ for the second. The implications of the change in exponents for the metrics of texts and for the origins of complex lexicons are analyzed.

Keywords: power laws, Zipf's law, word frequency, complex lexicons, language evolution

[1]Corresponding author.
Department of Physics, Universitat Politècnica de Catalunya (UPC)
Campus Nord, Mòdul B5, 2on pis, Despatx B5202
O8034 Barcelona, SPAIN
Phone: +34 93 4017056
FAX: +34 93 4017100
e-mail: ramon@complex.upc.es

1 Introduction

The Zipf's law for words, by G. K. Zipf (Zipf, 1972) , is one of the most fundamental and popular achievements of quantitative linguistics and the origin of a wide range of hypothesis about its origin (Li, WWW). Despite its apparent robustness (Naranan & Balasubrahmanyan, 1998; Balasubrahmanyan & Naranan, 1996), Zipf's law is an empirical observation and not a law in a rigourous sense (Casti, 1995; Li, 1998). In this context, Zipf's law has been assumed but not explained in recent models for the evolution of syntactic communication (Nowak, Plotkin, & Jansen, 2000) and is an obvious ingredient for any theory of language evolution.

Original Zipf's law[2] linked i, the rank of a word (in a list of words decreasingly ordered by frequency) with $P(i)$, its frequency. The relation follows a power law in the form:

$$P(i) = p_1 i^{-\alpha} \tag{1}$$

where $\alpha \approx 1$ (Zipf, 1972; Casti, 1995; Tsonis, Schultz, & Tsonis, 1997) and p_1 is the probability of the most frequent word.

The same law can also be presented as probability density function:

$$Q(j) \propto j^{-\beta} \tag{2}$$

where $Q(j)$ is the probability that a word is present j times in a text. Although both the rank distribution and the word frequency spectrum can be modeled in many ways

[2]G. K. Zipf discovered many rank-probability relations. Since we will focuse on that of words, we will simply hereafter refer to it as the Zipf's law.

(Tuldava, 1996; Balasubrahmanyan & Naranan, 1996; Naranan & Balasubrahmanyan, 1998), we adopt a power law for simplicty reasons.

We can relate the rank with the probability density function. Let us denote by $m_n = TQ(n)$ the number of words having population n, where T is the total number of word in the sample. Then, the rank is given by

$$R(n) = \int_n^\infty m_{n'} dn' \tag{3}$$

and the most frequent word has $R = 1$, the second most frequent word has $R = 2$, and so on, for decreasing values of n in the integral. Eq. 3 establishes a general relation between the rank of an event in the sample and the probability distribution according to the event frequency. Substituting $R \propto n^{-1/\alpha}$ (obtained from Eq. 1) and Eq. 2 in Eq. 3 we immediately get $n^{1-\beta} \simeq n^{-1/\alpha}$, from where

$$\alpha = \frac{1}{\beta - 1} \tag{4}$$

$$\beta = \frac{1}{\alpha} + 1 \tag{5}$$

If $\alpha = 1$ then β should be 2.

It can be observed in the plots of (Zipf, 1972; Casti, 1995; Tsonis et al., 1997) that the law provides a good fit for the smallest ranks (acknowledging some deviations at the very begining of the ordering discussed in (Tsonis et al., 1997; Li, 1998)) but no attention has been paid to the deviations in the tail. We will show that such deviations are much more important than expected.

2 Desagreements

One of the desirable properties of a law (as it happens with common physical laws) is to allow for accurate predictions.

The predicted number n of differents words of a text formed by T words, can be obtained by applying the Zipf's law and solving the following equation

$$\frac{1}{T} = p_1 n^{-\alpha} \tag{6}$$

where $1/T$ is the lowest probability that can be achieved by a word in a text of size T. From Eq. 6 we obtain

$$n = [Tp_1]^{1/\alpha} \approx Tp_1 \tag{7}$$

We processed [3] $T \approx 9 \cdot 10^7$ words of the British National Corpus (BNC) a corpus of modern English, both spoken (10%) and written (90%) [4]. We obtained $P(1) = 0.0601046$, $\alpha = 1$ (power law regression). Unfortunately, $n = 588,030$ was very far from $\hat{n} \equiv 5.6 \cdot 10^6$. The big deviation observed could be attributed to a poor statistics or a bad fitting of the parameters intervening in the prediction, p_1 and α. We will show that there is a deeper reason.

We computed the probability density function of the frequency (in number of occurences) of the BNC. More precisely, the probability $P(k)$ that a word occurs k times in the corpus. The left half of the plot, shown in Figure 2, revealed a well-defined

[3] Words different than proper noun were lowercased. Marks were excluded. Inflected forms of the same (root) word were treated as different words.

[4] BNC is a collection of text samples (generaly not longer than $45,000$ words). It is syncronic (it includes imaginative texts from 1960, informative texts from 1975), general (not specifically restricted to any particular subject field, register or genre), monolingual (it comprises text samples which are substantially the product of speakers of British English) and mixed (it contains both examples of both spoken and written English). Additional information is available at http://info.ox.ac.uk/bnc.

power law relationship between $Q(j)$ and j whose exponent was $\beta = 1.5$. The value obtained was 1.6, but removing the two first points, corresponding to the most uncommon words, and thus corresponding to the frequences being the most difficult to estimate, 1.5 was obtained (linear regression, $\beta = 1.52 \pm 0.008$). In contrast, Eq. 5 predicted $\beta = 2$. In addition, the plot of the probability density function in Figure 2 was specially clear. A question of bad statistics or fitting again?

FIGURE 1

3 Rethinking the law

A more carefull sight of the rank ordering plot on our data revealed the existence of two different exponents in the same rank ordering plot (Figure 1). $\alpha_1 = \alpha \approx 1$ and $\alpha_2 \approx 2$ seem appropiate for ranks $i < N \in (10^3, 10^4)$ and $i \geq N$, respectively. Thus, the frequency of words becomes a double law, the initial Zipf's law and a more sloping decay,

$$P(i) = \begin{cases} p_1 i^{-\alpha_1} & \text{if } i < N \\ N^{\alpha_2} p_N i^{-\alpha_2} & \text{otherwise} \end{cases} \tag{8}$$

where p_N is a the probability of the n-th most frequent word (it can also be obtained from Eq. 1 and thus be $1/p_1 N_1^{\alpha} \approx p_n/N$).

FIGURE 2

Let $x = [T p_1(1)]^{1/\alpha_1}$. According to 8 and being $1/T$ the smallest probability, the

number of different words predicted is

$$\hat{n} = \begin{cases} [Tp_1]^{1/\alpha_1} & \text{if } Tp_N < 1 \\ N\,[Tp_N]^{1/\alpha_2} & \text{otherwise} \end{cases} \tag{9}$$

where $p_{1,000} = 1.06292 \cdot 10^{-4}$, $p_{5,000} = 1.71864 \cdot 10^{-5}$ and $p_{6,000} = 1.34702 \cdot 10^{-5}$.

The value of $\hat{n}$ calculated through Eq. 9 is $213,570$, much closer to the real value. Figure 3 shows the value of n, $\hat{n}$, obtained through Eq. 7) and 9; $N = 6,000$) and Ebeling/Pöschel approximation (Ebeling & Pöschel, 1994) as a function of T.

FIGURE 3

4 Discussion

A single slope $\alpha = 1$ can only be attributed to a superficial look on small-sized texts in which deviations in the tail of the distribution (of the rank-ordering plot) were attributed to finite size effects instead of a different exponent. Many previous work on English was performed on relatively small texts, i.e. $260,430$ words (Zipf, 1972), $59,498$ words (Casti, 1995), $20,000$ words (Tsonis et al., 1997), far from the $\approx 9 \cdot 10^7$ words of the BNC we processed.

For long texts, the number of different words is mainly due to the second expression in Eq. 9. A relation $n \propto T^{-1/\alpha_2}$ was previously shown in (Ebeling & Pöschel, 1994) More precisely, $n = 22.8T^{0.46}$.

The two observed exponents divide words in two different sets: a kernel lexicon formed by $\approx N$ versatile words and an unlimited lexicon for specific communication.

We suggest that the size of the kernel lexicon is related with constrains of capacity of human brain. As a matter of fact, there is evidence of a relationship between characteristic size limitations and inflection points of power law exponents (Cancho & Solé, 2000).

The change of the exponent of the power law decay of the mutual information as a function of the distance between words agrees with the average length of sentences. We suggest that here the change in exponents is related with the average amount of words that human brain is efficiently able to store and use. Words with the highest rank are very specyfic and obviously not shared by all speakers. According to the intersection of the lines aproximating the two regimes of $P(i)$ in Figure 1, the kernel lexicon of the BNC would be formed by 5,000-6,000 words.

The existence of a kernel lexicon raises the question of how small can be a lexicon without drastically empoverishing communication. Pidgin languages provide examples of very small lexicons. Estimates of the number of items of a pidgin vary from about $300 - 1500$ words, depending on the language (Romaine, 1992, 1988). The number of lexical items of a speaker of an ordinary language is about $25,000 - 30,000$ (clearly not enough for the more than $500,000$ different words of the BNC) [5] while this amount is $1,500$ for a Tok Pisin speaker. It has been argued that these $1,500$ words can be combined into phrases so as to say anything that can be said in English (Hall, 1953). As

[5] Although lexicon size estimates very often rely on roughly approximated counts, the Waring-Herdan's recursive model for frequency spectrum allows to perform more accurate counts. This model straighforwardly allows for the calculation of the number of words which are known by an author that do not appear in the sample, m_0. If L is the number of different word in the sample, it has been shown that A.H. Tammsaare's lexicon contained (by the time the sample was written) about $L + m_0 = 8,228 + 25,147 = 33,000$ words. See (Tuldava, 1996) for more details.

expected, words of such small lexicons are very multifunctional and a circumlocution is often recurred for covering the lexicon gaps. The transition from the exponent α_1 to α_2 takes place in the interval of rank $10^3 < i < 10^4$. We suggest that common languages also have a lexicon of this kind, hidden by an unlimitited specific lexicon. Notice that although the size of the lexicon of a speaker can be very big, what counts for a successful communication are the words shared (stored and used) with the maximum number of speakers, that is, the words in the kernel lexicon.

The morpholocical simplicity and semantic generality that characterize pidgin and other known simplified lexicons (Romaine, 1992) respect to complex lexicons can also be indentified for the kernel lexicon. Table 1 summarizes them with examples from the BNC.

Some authors have pointed out the existence of two domains in the frequency of words (Naranan & Balasubrahmanyan, 1998), whose slopes agree with ours, or even three (Tuldava, 1996). Tuldava (Tuldava, 1996) determined three slopes for the rank distribution in the following ranges:

$$\begin{aligned} i = 1 - 30 \quad &- \quad \alpha_1 = 0.7 \\ i = 30 - 1,500 \quad &- \quad \alpha_2 = 1.1 \\ i = 1,500 \quad &- \quad \alpha_3 = 1.4 \end{aligned}$$

Statistics were performed on A. H. Tammsaare's novel "Truth and Justics" and only lexemes were considered. The transition between the 2^{nd} and the 3^{rd} regime takes place

in a rank closer to pidgin lexicons size. Inflected forms of the same word were counted as different words in our statistics, suggesting that the rank at which the change in exponents takes place could be reduced. The slope of the less frequent words regime (1.4) is remarkably different than BNC's (2). Further study is needed for determining the origin of this disagreement.

We calculated the proportion of words of a text belonging to the kernel lexicon as a function of N, $S(N) = \sum_{i=1}^{N} P(i)$, being $P(i)$ the real probability of the i-th word) in order to illustrate the importance of the kernel. $S(1,000) = 0.69$, $S(4,000) = 0.84$, $S(5,000) = 0.86$ and $S(6,000) = 0.87$ show how recurring are such words. An exception to the universality of the frequency exponents is Shakespeare, in which $\beta = (-)1.6$. It is said that Shakespeare used one half of the words extant at that time. In a sample of 884.647 words, he used more than $30,000$ different words, of which at least 8% were of his own creation (Balasubrahmanyan & Naranan, 1996) [6]. We suggest that he broke the commones of the kernel words in the seek of wide and dazzling vocabulary. The big amount of words of his own creation supports it. To sum up, the two frequency domains separate two clearly distinguishable classes of words.

TABLE 1

[6] Although only content words were considered, the absence of a domain corresponding to a kernel lexicon can not be attributed the absence of content words because such domain usually contains them.

Acknowledgements

Authors want to thank Susanna Manrubia for helpful discussions and valuable support.

This work was supported by grants of the Generalitat de Catalunya (FI/2000-00393, RFC) and the CICYT (PB97-0693, RVS) and the Santa Fe Institute (RVS).

References

Balasubrahmanyan, V. K., & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*(3), 177-228.

Cancho, R. F. i, & Solé, R. V. (2000). Long-range correlations and characteristic size in human language. *To be submitted to Europhysics Letters.*

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

Ebeling, W., & Pöschel, T. (1994). Entropy and long-range correlations in literary english. *Europhysics Letters*, *26*(4), 241-246.

Hall, R. A. (1953). *Haitian creole: Grammar, texts, vocabulary.* American Folcklore Society Memoire.

Li, W. (1998). Letters to the editor. *Complexity*, *3*, 9-10. (Comments to "Zipf's Law and the structure and evolution of languages" A.A. Tsonis, C. Schultz, P.A. Tsonis, COMPLEXITY, 2(5). 12-13 (1997))

Li, W. (WWW). *Zipf's law.* http://linkage.rockefeller.edu/wli/zipf.

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

Nowak, M. A., Plotkin, J. B., & Jansen, V. A. (2000). The evolution of syntactic communication. *Nature*, *404*, 495-498.

Romaine, S. (1988). *Pidgin and creole languages.* London: Longman.

Romaine, S. (1992). The evolution of linguistic complexity in pidgin and creole languages. In J. A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (p. 213-238). Addison Wesley.

Tsonis, A. A., Schultz, C., & Tsonis, P. A. (1997). Zipf's law and the structure and evolution of language. *Complexity*, *3*(5).

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, *3*(1), 38-50.

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

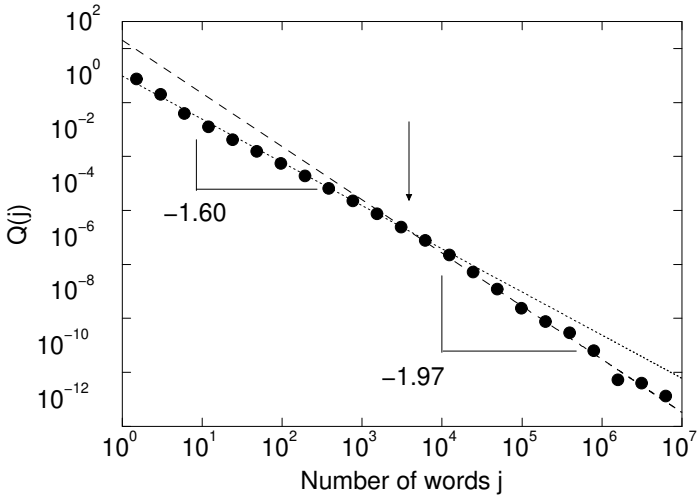

Figure 1: Probability that a word occurs i times. The first and the second power law decays have exponent $\alpha_1 = 1.06 \pm 0.04$ and $\alpha_2 = 1.97 \pm 0.06$, respectively ($r > 0.99$ in both cases). Statistics on the BNC ($T \sim 9 \cdot 10^7$ words, $n \sim 588,030$)

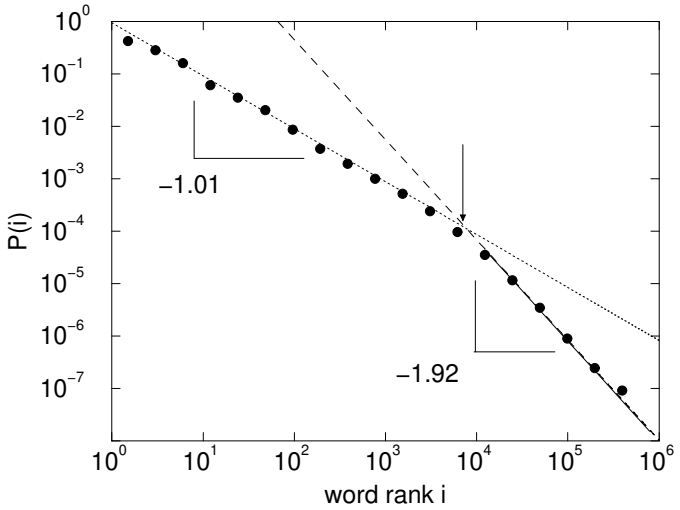

Figure 2: Probability of a word as a function of its rank i, $P(i)$. The first and the second power law decays have exponent $\alpha_1 = 1.01 \pm 0.02$ and $\alpha_2 = 1.92 \pm 0.07$, respectively ($r > 0.99$ in both cases). Statistics on the BNC ($T \sim 9 \cdot 10^7$ words, $n = 588,030$)

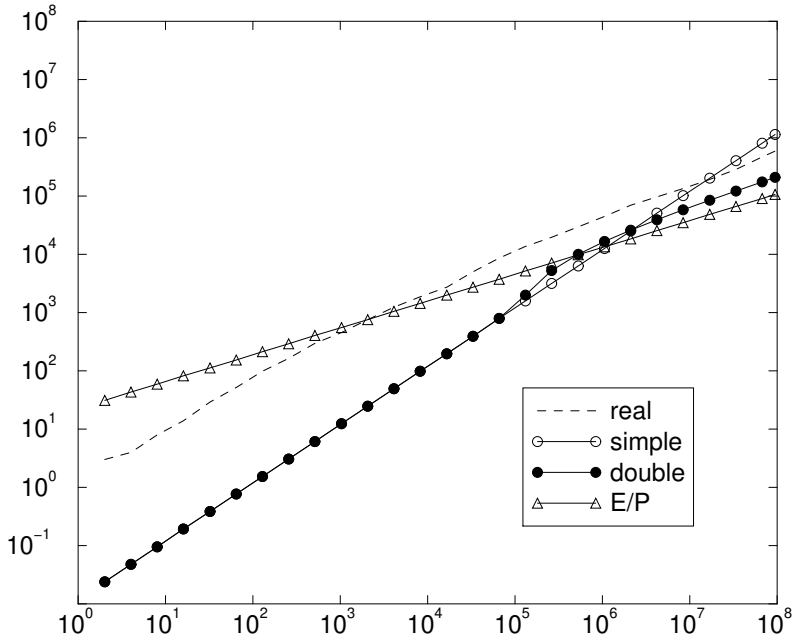

Figure 3: Number of different words as a function of the total number of words , T, of the sample. The real number is accompanied by estimations performed with the Zipf's law (Eq. 7), the two regime frequency observation (Eq. 9; $N = 6,000$) and the Ebeling/Pöschel approximation. .

	kernel lexicon	rest of the lexicon
generalyty of terms	generic terms rather than specific terms (e.g. is_{9}, see_{96}, $group_{233}$, $live_{634}$, $know_{1,435}$ and $bird_{1,981}$)	larger vocabulary in a given domain (e.g. $biplane_{39,903}$, $coda_{43,482}$, $scarps_{68,727}$, $mycelium_{111,889}$, $anticoagulants_{113,286}$ and $microscopium_{432,607}$)
complexity of words	monomorphemic words [7] (e.g. it_{7}, $made_{104}$, $year_{120}$, $hand_{246}$ and $mad_{3,312}$)	compounds (e.g. $airbrakes_{35,182}$, $fingerpriting_{53,988}$, $peachtree_{137,080}$, $breakdance_{163,284}$, $fingerlocks_{439,217}$ and $spillway_{453,615}$) and morphologically complex words (e.g. $childishly_{46,541}$, $literariness_{55,355}$, $thoughtlessness_{65,489}$, $overindebtedness_{97,885}$, $proletarianized_{103,707}$ and $multiculturated_{437,580}$)

Table 1: Comparison between the kernel lexicon and the rest of the lexion. The intervening features were originally devised for comparing simple lexicons (pidgin,creole,...) and complex lexicons. Example words are subindexed by its rank.