Eur. Phys. J. B **44**, 249–257 (2005)
DOI: 10.1140/epjb/e2005-00121-8

THE EUROPEAN PHYSICAL JOURNAL B

The variation of Zipf's law in human language

R. Ferrer i Cancho[a]

INFM udR Roma1, Dip. di Fisica. Università "La Sapienza". Piazzale A. Moro 5. 00185 Roma, Italy

Received 20 August 2004
Published online 20 April 2005 –

Abstract. Words in humans follow the so-called Zipf's law. More precisely, the word frequency spectrum follows a power function, whose typical exponent is $\beta \approx 2$, but significant variations are found. We hypothesize that the full range of variation reflects our ability to balance the goal of communication, i.e. maximizing the information transfer and the cost of communication, imposed by the limitations of the human brain. We show that the higher the importance of satisfying the goal of communication, the higher the exponent. Here, assuming that words are used according to their meaning we explain why variation in β should be limited to a particular domain. From the one hand, we explain a non-trivial lower bound at about $\beta = 1.6$ for communication systems neglecting the goal of the communication. From the other hand, we find a sudden divergence of β if a certain critical balance is crossed. At the same time a sharp transition to maximum information transfer and unfortunately, maximum communication cost, is found. Consistently with the upper bound of real exponents, the maximum finite value predicted is about $\beta = 2.4$. It is convenient for human language not to cross the transition and remain in a domain where maximum information transfer is high but at a reasonable cost. Therefore, only a particular range of exponents should be found in human speakers. The exponent β contains information about the balance between cost and communicative efficiency.

PACS. 87.10.+e General theory and mathematical aspects – 89.75.Da Systems obeying scaling laws

1 Introduction

Word frequencies in human language arrange themselves according to the so-called Zipf's law. If $P(f)$ is the proportion of words whose frequency is f in a given sample (e.g. a text), we say the sample follows Zipf's law [1,2] if

$$P(f) \sim f^{-\beta} \tag{1}$$

where $\beta > 0$.

Although different functions have been proposed for modeling word frequencies [3,4], the fundamental trend described by equation 1 seems to have no known exception in humans. As far as we know, Zipf's law is a universal property of world languages [5]. We have typically $\beta \approx 2$ in word frequencies of single author samples [6–9], but significant deviations from that value have been reported:

- $\beta > 2$ in fragmented discourse schizophrenia. The speech is characterized by multiple topics and the absence of consistent subject. The lexicon of such a text may be varied and chaotic [10,11]. $\beta \in [2.11, 2.42]$ is found. That type of schizophrenia is found in early stages of the disease but not all early stages follow that pattern.
- $1 < \beta < 2$ in advanced forms of schizophrenia [1,10,11]. Texts are filled mainly with words and word combinations related to the patients obsessional topic. The variety of lexical units employed here is restricted and repetitions are many. $\beta = 1.6$ is reported in [10,11].
- $\beta = 1.6$ in very young children [12,10]. Older children conform to the typical $\beta \approx 2$ [13].
- $\beta = 1.7$ in military combat texts [11,14].
- $\beta = 3.35$ for nouns in multiauthor collections of texts [15]. Smaller values of β but still above the typical $\beta = 2$ have been found in nouns from single author samples. More precisely, $\beta \in [2.15, 2, 32]$ [5].
- Exponents larger than $\beta \approx 2$ can be obtained as a result of deficient sampling from a text with the typical $\beta \approx 2$ [10,11].

Therefore, exponents seem to be constrained to a very particular domain. More precisely, $\beta \in [1.6, 2.42]$, excluding $\beta = 3.35$ for nouns because it is restricted to a particular word type. Is such a particular domain a mere consequence of our limited knowledge or is there a theoretical explanation going beyond our limited capacity for studying all the possible atypical forms of language? Is there a barrier preventing speakers from crossing the lower and upper bounds of the interval of empirical values of β? The remainder of the paper is devoted to shed light on the

[a] e-mail: ramon@pil.phys.uniroma1.it

previous questions. Notice that there could be a trivial explanation for the variation of β, i.e. random fluctuations around a mean $\beta = 2$. Although that could be a reasonable hypothesis for the variation in normal adult speakers, the problem is that schizophrenics, children and military combat texts do not seem to fill the continuum from $\beta = 2$ to $\beta \approx 1.6$. There seems to be a gap. Furthermore, it is not clear why variation should would be approximately constrained to the interval $[1.6, 2.4]$. Again, our limited knowledge or publications bias (i.e. the tendency to report only the clearly anomalous cases) could be showing ghost gaps and barriers. At this point, two ways can be taken: performing studies in larger sets or gaining theoretical knowledge about the variation of Zipf's law. The latter is the aim of the present paper.

We assume that T is the length of a text in words (so $f \leq T$). In order to understand the depth of those questions, it is important to bear on mind that the value of β when $T \to \infty$ has consequences based only on equation (1). More precisely, $\beta > 1$ is needed if $P(f)$ must be a probability function, $\beta > 2$ if f must have finite mean and $\beta > 3$ if f must have finite variance [16]. The minimum and maximum real value of β are far from the threshold exponents in every requirement, suggesting there may be a further reason. We need further information than word frequencies in order to constraint the range of possible values. We will show that taking into account that words have meaning is the key.

Atypical exponents are the Achilles heel of the null hypothesis and models that have been propose for Zipf's law. Here we focus on two null hypothesis: intermittent silence and Simon's model. By null hypothesis, we mean processes lacking some fundamental aspect of why and how human words are used. For instance, intermittent silence and Simon's model neglect that words are used according to their meaning. The more meanings a real word has, the higher its frequency of use [17–19]. The distinction between null hypothesis and model in the strict sense is not tight. What some researchers may call a null hypothesis due to its extreme simplification of the real problem, may be called a model in the strict sense by researchers considering those simplifications mere consequences of a high level of abstraction. In other words, what some researchers may consider neglectable, can be unneglectable for others. Some researchers may consider intermittent silence or Simon's model not even suitable hypothesis for Zipf's law. Specially for the latter researchers, it is important to bear on mind that the believe that Zipf's law in human words can be explained by intermittent silence is widespread in science [20–24], so the present discussion is well-motivated. The terms null hypothesis and model should be read with a flexibility depending on the point of view one is taking. The aim of what follows is not to reach an agreement about how intermittent silence and Simon's model should be classified, but to call the attention to the fact that those models can not explain the variations in the exponent reported above.

Intermittent silence consists of generating a random sequence of characters including letters (or phonemes) and blank spaces (or silences) [24–28]. A words is defined as a sequence of letters between blanks. Intermittent silence can not consistently explain the values of $1 < \beta < 2$ in schizophrenics where β is clearly far from 2 [1]. If L is the set of letters, the exponent of an intermittent silence model as in [24] is

$$\beta_I = \frac{\log |L|}{\log(|L| + 1)} + 1, \tag{2}$$

where $|L|$ is the cardinality of L. It follows from equation 2 that $1 < \beta_I < 2$ ($|L| \geq 1$ is assumed) and $\beta_I < 1.9$ implies $|L| \leq 5$. Therefore, such disease implies a repertoire of letters (or phonemes) radically smaller from that of regular speakers, which is absolutely false. Simon's model is based on reusing the most frequently used words. The model adds a word to a text at each iteration. With probability ψ, the word added is such that has not appeared before in the text. With probability $1-\psi$, the word added is chosen at random from all the words in the text. The distribution of the process follows equation (1) with [29,30] exponent

$$\beta_S = 1 + \frac{1}{1-\psi}. \tag{3}$$

We have $\beta_S > 2$ for $\psi > 0$ [29], so Simon's model can not directly account for the word frequency distribution in all cases of schizophrenia.

A theory of word frequencies and human communication implies answering to four different questions:

1. Why words follow arrange according to Zipf's law (Eq. (1)).
2. Why humans typically choose $\beta \approx 2$?
3. Why is there variation in β?
4. What are the limits of that variation?

Many answers have been proposed for Questions 1-2 [1,4,5,8,15,24–27,29,31–44]. As far as we know, Questions 1 and 2 have only been answered assuming that words are used according to their meaning in [38] and [15], respectively. Choosing $\beta \approx 2$ could be the optimal solution for a conflict between hearer and speaker needs [38]. The present paper is focused on Questions 3 and 4.

In the present paper we provide an explanation for the limits of the variation in β which is consistent with the real range of exponents. The models starts with a mathematical formalization of the goal and the constraints of communication and a very basic assumption: words are used according to their meaning.

2 The model

We assume a general communication system mapping signals to stimuli. We have a set of n signals $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of m stimuli $R = \{r_1, ..., r_j, ..., r_m\}$. Here we assume that signals are approximately equivalent to words and stimuli are the basic ingredient for constructing word meaning. We assume that signals connect to stimuli and that connections are defined

by a binary $n \times m$ matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if s_i and r_j are linked and $a_{ij} = 0$ otherwise. What we mean here by stimuli can be explained in more detail. Different types of experiments have shown that words are associated to the activation of different brain areas [45]. Generally speaking, nouns tend to activate visual areas. Verbs tend to activate motor areas if the corresponding action can be performed by the individual and visual areas otherwise. The activated areas are associated to different types of stimuli experienced with the word. Let us take one of the definitions of the Webster's Revised Unabridged Dictionary (1913)[1] for the word 'write': "to inscribe on any material by a suitable instrument". In our view, the verb 'write' is associated to the motor stimuli of the action of writing and the visual (tactile, olfactive,...) stimuli of the instruments used for writing. The construction of a complex meaning would involve a structure combining different stimuli. From that point of view, a words in S does not refer to stimuli in R, but is merely associated to them. We do not claim that words in S refer to stimuli in R via A although they can. We do not use the term reference because it is stronger than association. In our example, 'write' can only refer to the motor stimuli of the action. 'write' can not be used for referring to the instrument used for writing although it is associated to it. The action and the instrument are both stimuli involved in the construction of the complex meaning of the verb 'write'. Defining word meaning is an open problem in different fields ranging from cognitive science to philosophy. In our view, complex meaning would emerge from the interaction between different stimuli. Referential associations are a subset of the associations defined by A. It makes sense to assume that the more stimuli a word is associated to, the highest the probability of using that word. It is important to notice that when we say a word has no meaning we are usually saying that it has not referential power but that does not imply that word has not associations with stimuli. Our framework is not inconsistent with the existence of words with no apparent meaning, such as prepositions, conjunctions or articles. Real words with no apparent meaning are the words with the highest frequencies. The five most frequent word in the British National Corpus[2], a large collection of text samples, are 'the, 'of', 'and', 'to' and 'a'. The framework here predicts that the most frequent words would have the largest amount of connections with stimuli in R. Since those connections are merely associative and sometimes referential there is no inconsistency here. Furthermore, that high amount of associations may underly the absence of referential power or meaning of that words. The uncertainty associated to the interpretation of highly connected words is so large [46] that reference can not be effectively attributed. Words with no meaning may have two different origins: words that have no links and words having too many links. It makes sense to think that words with no meaning may have an excess of connections rather than a lack of it, although those connections could be very weak due to the high frequency of the words involved [47].

[1] www.dict.org
[2] www.natcorp.ox.ac.uk

We define the probability that s_i a r_j are associated by the communication system as

$$p(s_i, r_j) = \frac{a_{ij}}{||A||}, \tag{4}$$

where $||A||$ is a normalization factor defined as

$$||A|| = \sum_{k=1}^{n} \mu_k \tag{5}$$

and $\mu_i = \sum_{k=1}^{m} a_{ik}$ is the number of stimuli of s_i as in [15]. Replacing equation (4) into $p(s_i) = \sum_{j=1}^{m} p(s_i, r_j)$ we get

$$p(s_i) = \frac{\mu_i}{||A||}. \tag{6}$$

Similarly, replacing equation (4) into $p(r_j) = \sum_{k=1}^{m} p(s_k, r_j)$ we get

$$p(r_j) = \frac{\omega_j}{||A||}, \tag{7}$$

where $\omega_j = \sum_{k=1}^{m} a_{kj}$ is the number of signals of r_j. Notice that here $p(r_j)$ is not independent of A as in [38].

Equations (6) and (7) contain our basic assumption that words are used according to their stimuli. More precisely, equation (6) states that a word is used with a frequency that is proportional to its number of stimuli. The choice of that extremely simple equation is supported by that the following points:

- Various models are capable of reproducing Zipf's with that assumption [15,46].
- Word frequency and number of meanings are positively correlated. It is reasonable to think that our stimuli here are also positively correlated with word frequency [17–19].
- Since we do not know the real relationship between word frequency and the stimuli defined here, we choose proportionality for simplicity reasons.

If $P(k)$ is the proportion of signals having k links we may write equation (6) as

$$p(s_i|\mu_i = k) = \frac{k}{n \langle k \rangle_P}, \tag{8}$$

where n is the number of signals and $\langle ... \rangle_P$ is the expectation operator over $P = \{P(1), ..., P(k), ...P(m)\}$. Following equation (4), the entropy (or uncertainty) associated to the interpretation of s_i is $H(R|s_i) = \log k$ if $\mu_i = k$ [15]. Thus, $H(R|S)$, the average entropy associated to the interpretation of a signal [48], defined as

$$H(R|S) = \sum_{i=1}^{n} p(s_i) H(R|s_i) \tag{9}$$

becomes

$$H(R|S) = \frac{\langle k \log k \rangle_P}{\langle k \rangle_P}. \tag{10}$$

The symmetric, $H(S|R)$ becomes

$$H(S|R) = \frac{\langle k \log k \rangle_Q}{\langle k \rangle_Q}, \tag{11}$$

where $Q = \{Q(0), ..., Q(k), ..., Q(n)\}$ and $Q(k)$ is the proportion of stimuli with k links. We are assuming that $Q(k)$ is defined for $k = 0$ while $P(k)$ does not because we allow unlinked stimuli but do not allow unlinked signals. Here we take the simplest distribution for Q, that is

$$Q \sim binomial\left(\frac{\langle k \rangle_Q}{m}, n\right). \tag{12}$$

Similarly, the signal entropy

$$H(S) = -\sum_{i=1}^{n} p(s_i) \log p(s_i) \tag{13}$$

can be written as

$$H(S) = \log(n \langle k \rangle_P) - \frac{\langle k \log k \rangle_P}{\langle k \rangle_P}. \tag{14}$$

Merging equation (10) and (14) we get

$$H(S) = \log(n \langle k \rangle_P) - H(R|S). \tag{15}$$

We define a function Ω that any biological communication system must minimize. The function is a combination of the goal of communication, that is, maximizing the information transfer [48] between the set of signals and the set of stimuli, $I(S, R)$, and the constrains imposed by the biology of the communication system, which tend to minimize $H(S)$, the entropy associated to signals [38]. $H(S)$ is the cost of the communication. Taking a linear combination for simplicity, we define Ω, the energy of a communication system [38], as

$$\Omega(\lambda) = -\lambda I(S, R) + (1 - \lambda) H(S), \tag{16}$$

where $0 \leq \lambda \leq 1$. λ is a parameter controlling the balance between $-I(S, R)$ and $H(S)$. When $\lambda = 0$ the goal of communication does not matter at all and the same happens to the constraints of communication when $\lambda = 1$.

If $p(s_i) \sim \mu_i$ (Eq. (6)) and Zipf's law (Eq. (1)) are assumed, it follows that $P(k)$, the proportion of signals with k links obeys

$$P(k) \sim k^{-\beta}. \tag{17}$$

We assume a fixed $P(k)$ or $P(f)$ given the surprising tendency of human language to arrange according to Zipf's law even in the atypical cases. Although there is variation in β, equation (1) has no known exceptions. Various models show that if the constraint of equation (17) is removed and $\Omega(\lambda)$ is minimized, Zipf's law is recovered for $\lambda = \lambda^*$, where λ is a critical value of λ [38,46].

We are interested in studying β^*, the value of β obtained when minimizing $\Omega(\lambda)$ for different values of λ while n and m are kept constant. Applied in the inverse sense, that is also a way of discovering λ given the value of β of a particular communication system (when n and m are known and constant). We assume that β^* is a single valued function. The numerical calculations in Section 3 make sure that there is a single value of β minimizing Ω. We may write equation (16) as

$$\Omega(\lambda) = (1 - 2\lambda) H(S) + \lambda H(S|R) \tag{18}$$

using $I(S, R) = H(S) - H(S|R)$, where $H(S|R)$ is the average entropy associated to choosing a certain signal for a certain stimulus [48]. Trivially, $\Omega(\lambda)$ minimizes $H(S|R)$ if $\lambda > 0$. Notice that, simultaneously, $H(S)$ is minimized if $\lambda < 1/2$ and maximized if $\lambda > 1/2$. In other words, the cost of communication is minimized for $\lambda < 1/2$ and maximized for $\lambda > 1/2$. That cost is irrelevant for $\lambda = 1/2$.

Notice that the outcome of minimizing $\Omega(\lambda)$ when $\lambda = 1/2$ is every stimulus with exactly one non-necessarily distinctive signal (if $n \geq m$). When $\lambda > 1/2$ the behaviour of the system should be similar, but taking into account that maximizing $H(S)$ implies that the signal linked to each stimulus should be distinctive because the $H(R|S)$ inside $H(S)$ (recall Eq. (15)) is being implicitly minimized. Thus, it is easy to see that equation (17) (or equivalently Eq. (1)) with $\beta \to \infty$ would hold when $\lambda > 1/2$. Since as far as we know real human speech shows finite β, the interesting values of λ are the large ones that still satisfy $\lambda < 1/2$. We will show that finite values of β are only found in the range $0 < \lambda < 1/2$. Furthermore, we will show that values are small.

3 Results

We study β^*, the value of β in equation (17) minimizing $\Omega(\lambda)$ versus λ for different values of n (the number of signals) and m (the number of stimuli).

$\Omega(\lambda)$ is minimized numerically by exploring exhaustively the interval $[0, \beta_{max}]$ where $\beta_{max} = 10$ with a resolution $\epsilon = 10^{-2}$ for β and λ. We have seen that β^* should diverge beyond $\lambda = 1/2$ at the very latest. Since $\Omega(0)$ gives a minimum in $(0, \beta_{max})$ (we assume that is the global minimum), β^* must diverge at some λ between 0 and 1/2. Therefore, we need a method to distinguish true divergences from the limitations of exploring the interval $[0, \beta_{max}]$. We define $\beta^*(\lambda)$ as the value of β^* as a function of λ (when n and m do not change). Thus, we explore λ from 0 to 1 (the order of exploration is important) and whenever the value of β minimizing $\Omega(\lambda)$ is β_{max}, we conclude $\beta^* \to \infty$ only if one of the following conditions is true:

1. $\beta^*(\lambda - \epsilon) \ll \beta_{max}$. That condition is supported by the a priori expectation that β^* must jump at a certain value of $\lambda \leq 1/2$ from a finite value to infinity.
2. $\beta^*(\lambda - \epsilon)$ diverges according to the condition 1 or 2.

Ω is calculated assuming the total number of links in A is $||A|| = n \langle k \rangle_P$ in equation (15). That is an approximation because the right side of the previous equation is a continuous value while the left side is a discrete one. The

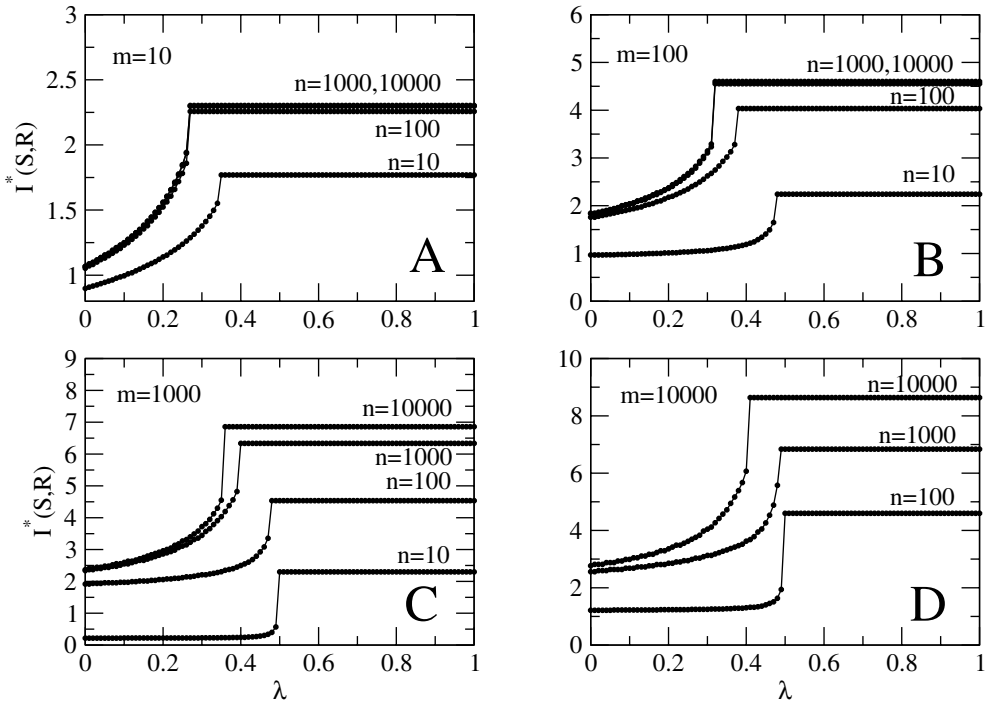

Fig. 1. $I^*(S,R)$ versus λ for $m = 10$ (A), $m = 10^2$ (B), $m = 10^3$ (C) and $m = 10^4$ (D). $I^*(S,R)$ is the information transfer for $\beta = \beta^*$, β^* is the value of β minimizing $\Omega(\lambda)$, β is the exponent of Zipf's law, Ω is the energy function that communication minimizes, n is the number of signals and m is the number of stimuli. λ tunes the balance between communicative efficiency and the cost of communication. When $\lambda = 0$ communication is totally balanced towards saving the cost of communication whereas when $\lambda = 1$ is totally balanced towards the communicative efficiency. Natural logarithms are used for showing $I(S,R)$. The series for $m = 10000$ and $n = 10$ are not shown due to the limitations of the fast calculations used.

approximation is useful in order to run calculations faster. Nonetheless, that approximation may lead to inconsistent results. That is why the series for $m = 10$ and $n = 10000$ are not shown in any of the following figures.

The value of $I^*(S,R)$ undergoes a sudden transition to maximum $I(S,R)$ when $\lambda = \lambda^*$ is crossed (Fig. 1). Interestingly, the cost of the communication, $H(S)$ also becomes maximum after the transition (Fig. 2). The sharpness of the changes when $\lambda = \lambda^*$ suggests a phase transition to maximum $I(S,R)$ and maximum $H(S)$ as in the model in [38].

β^* grows with λ, but it does not behave gradually in the whole domain (Fig. 3). There is a critical value of λ, λ^*, beyond which β^* diverges. That divergence is non-trivial since it does not necessarily happen when $\lambda^* \geq 1/2$.

Increasing m tends to decrease the maximum finite value of β^* $\beta \approx 2.41$ (Fig. 4). Interestingly, the maximum finite value of β^* reaches the apparent upper bound $\beta \approx 2.41$ found in real exponents. Furthermore, the minimum value of β^* is given by $\lambda = 0$. $\beta > 1.5$ is obtained in Figure 5 for sufficiently large m, suggesting there is a minimum value for β^* even when maximizing $I(S,R)$ is neglected (that is, when $\lambda = 0$). Here, β is explored with a resolution of 10^{-3}. Actually, the value of β minimizing $\Omega(0)$ grows with m (Fig. 5) but very slowly. Notice that β^* is independent of n in that case since n is kept constant during the minimization process.

Neglecting the fact that signals are linked to stimuli leads to totally different results. Now we assume $p(s_i) > p(s_{i+1})$ (for $1 \leq i < n$) without loss of generality. If

$$p(s_i) \sim i^{-\alpha} \tag{19}$$

then it follows [3,7] equation (1) with

$$\beta = 1/\alpha + 1. \tag{20}$$

Thus, we calculate α^*, the value of α minimizing $H(S)$ and obtain β^* using equation (20). Now, β^* depends on n, the number of signals. R is irrelevant here. $H(S)$ is minimized if and only if $p(s_i) = 1$ if $i = 1$ and $p(s_i) = 0$ otherwise. That configuration is obtained by equation (19) when $\alpha \to \infty$, so using equation (20) we get

$$\beta^* = \lim_{\alpha \to \infty} \frac{1}{\alpha} + 1 = 1 \tag{21}$$

$\beta^* = 1$ is very different from the $\beta^* > 1.5$ obtained when stimuli are relevant (Fig. 5). Neglecting that words have meaning can not explain the minimum value of $\beta \approx 1.6$ found in human language.

4 Discussion

Here we have assumed Zipf's law and have studied the effect of tuning the balance between the goal ($I(S,R)$)

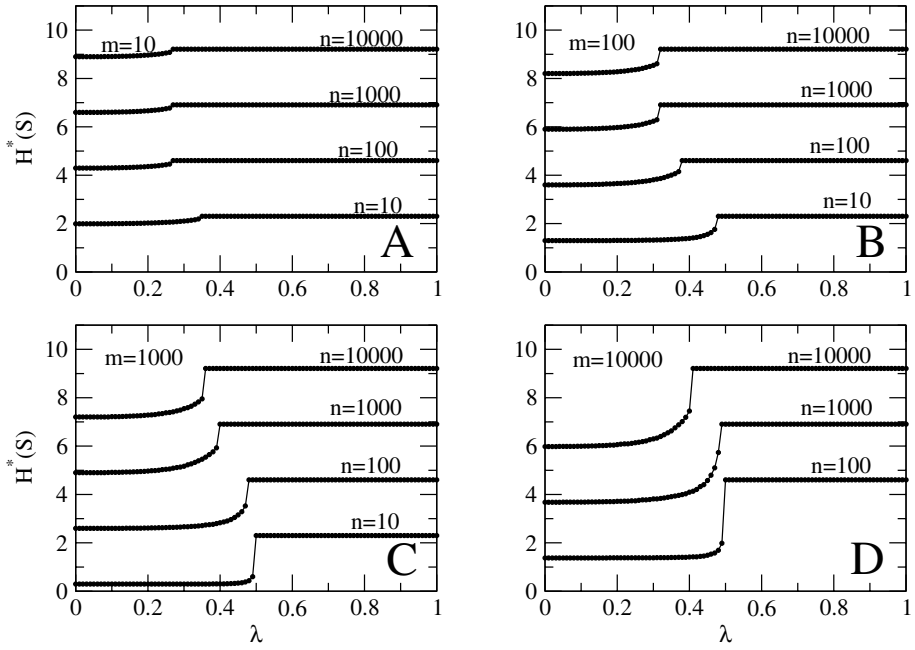

Fig. 2. $H^*(S)$ versus λ for $m = 10$ (A), $m = 10^2$ (B), $m = 10^3$ (C) and $m = 10^4$ (D). $H^*(S)$ is the signal entropy for $\beta = \beta^*$, β^* is the value of β minimizing $\Omega(\lambda)$, β is the exponent of Zipf's law, Ω is the energy function that communication minimizes, n is the number of signals and m is the number of stimuli. λ tunes the balance between communicative efficiency and the cost of communication. When $\lambda = 0$ communication is totally balanced towards saving the cost of communication whereas when $\lambda = 1$ is totally balanced towards the communicative efficiency. Natural logarithms are used for showing $H(S)$. The series for $m = 10\,000$ and $n = 10$ are not shown due to the limitations of the fast calculations used.

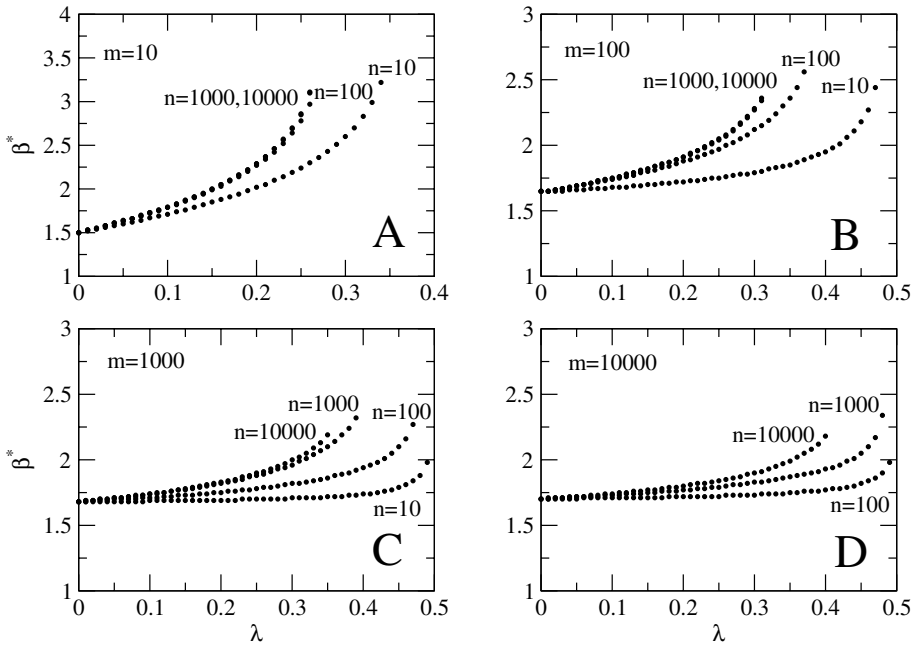

Fig. 3. β^*, the value of β minimizing $\Omega(\lambda)$ for $m = 10$ (A), $m = 10^2$ (B), $m = 10^3$ (C) and $m = 10^4$ (D). β is the exponent of Zipf's law, Ω is the energy function that communication minimizes, n is the number of signals and m is the number of stimuli. λ tunes the balance between communicative efficiency and the cost of communication. When $\lambda = 0$ communication is totally balanced towards saving the cost of communication whereas when $\lambda = 1$ is totally balanced towards the communicative efficiency. The series for $m = 10000$ and $n = 10$ are not shown due to the limitations of the fast calculations used.

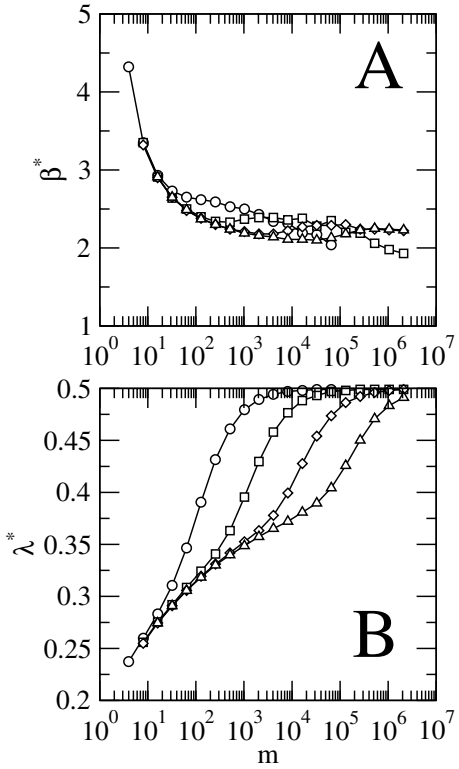

Fig. 4. A. The maximum finite value of β^* versus m for different values of n: $n = 10^2$ (circles), $n = 10^3$ (squares), $n = 10^4$ (diamonds) and $n = 10^5$ (triangles). β^* is the value of β minimizing $\Omega(\lambda)$ and $\lambda \in [0, 1]$. B. The same for λ^*, the corresponding value of λ for the values of β^* shown in A.

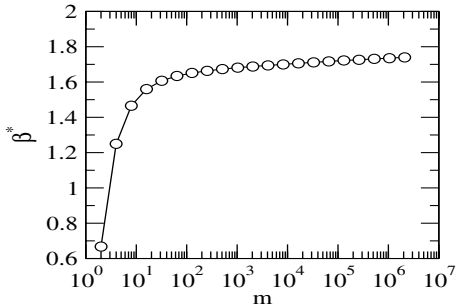

Fig. 5. β^*, the value of β minimizing $\Omega(0) = H(S)$ versus m.

and the cost ($H(S)$) of communication. We have used the parameter λ for controlling that balance. $\lambda = 0$ implies neglecting the goal of communication in the model. It is important to notice that the communicative efficiency is not absolutely destroyed when $\lambda = 0$. Zipf's law from equation (17) means that some signals have a few links even when $\beta < 2$, which favours implicitly the goal of communication. The model in [15] suggests that the assumption of Zipf's law (Eq. (17)) could be the outcome of implicitly favouring the goal of communication. Recall $P(k)$ and $Q(k)$ are, respectively, the proportion of signals and stimuli having k links. That model explains why scaling would be expected if the entropy of $P = \{P(k)\}$ is free and maximized, $Q = \{Q(k)\}$ is also free and

$$\sum_{i=1}^{n} H(R|s_i) = \langle \log k \rangle_P \tag{22}$$

is constrained. The latter constraint is similar to the goal of communication used here. Here maximizing $I(S, R) = H(R) - H(R|S)$ implies minimizing $H(R|S)$ which is turn similar to constraining $\langle \log k \rangle_P$ (Eq. (22)) to a small value.

Figure 5 shows there is a non-trivial lower bound for β^*. Minimizing $\Omega(0) = H(S) = \log(n \langle k \rangle_P) - H(R|S)$ implies a conflict between minimizing $\langle k \rangle_P$ and $H(R|S)$ that can not be resolved by and extremal value of β. That may explain why real exponents do not cross the apparent $\beta \approx 1.6$ barrier (for sufficiently large m).

Two groups of schizophrenic patients with different symptoms and different values of β have been found. The growth of β^* versus λ in Figure 3 suggests the group with $\beta < 2$ is constrained by the cost of communication more than normal speakers. Besides, the group with $\beta > 2$ could be favouring the goal of communication (here the word 'favouring' does not imply a voluntary action) more than normal speakers. It has been suggested that the notion of the self, the distinction between speaker and hearer, and more particularly the distinction between the signals that the individual generates as speaker and those that he receives as hearer, is a fundamental component of schizophrenia [49]. The latter is crucial since without a proper notion of the self and a proper identification of the receiver, communication may be balanced towards saving the cost of communication as much as possible, which could be the case of schizophrenic patients with $\beta < 2$. To sum up, schizophrenics with $\beta < 2$ would be saving as much cost of the communication as possible where as schizophrenics with $\beta > 2$ would be paying too much.

Child speech acts reported in [12] share $\beta \approx 1.6$ with schizophrenic speech with $\beta < 2$. The coincidence may not be by chance. In both cases the model suggests the cost of communication can not be overcome as in normal adult speech. Obviously, the reasons are different. Children unaware of the communicative value of their speech or constrained by their brain limitations in a more fundamental way would balance communication towards saving the cost communication as much as possible. Whereas insufficient development is the reason of that particular balance in children, the loss of the perspective of the hearer could be the reason in that type of schizophrenic patients.

The fact that military combat texts and some children and schizophrenics have similar exponents suggests that those texts could be shaped by an atypical balance saving the cost of communication, or equivalently from equation (16), dissociated from the goal of communication more than normal adult speakers. The possibility that the actual mechanism is the same as in children or schizophrenic patients can not be denied. No matter how different could be the mechanisms arranging word frequencies in children, schizophrenic patients (with $\beta < 2$) and military combat texts, all three systems seem to obey a common principle: minimizing $H(S)$.

We have seen that tuning the balance between the goal of communication and the cost of communication may lead to a non-gradual behaviour of β^*, $I^*(S,R)$ and $H^*(S)$. Interestingly, increasing λ beyond λ^* has no effect. The model presented here explains why human language maximizes the communication transfer, although $H(S)$ is not maximum and β is finite.

The model shows that it is not necessary to have $\lambda = 1$ for having the maximum $H(S)$, the cost of communication. Furthermore, maximum $H(S)$ is always obtained when $\lambda > 1/2$. It is convenient for language to favour the constraints of communication over the goal of communication in order to escape from paying the maximum $H(S)$. Therefore, the real values of λ should be slightly below λ^*. Since $\lambda^* < 1/2$, it follows from the model here that the constraints of communication must be stronger than the goal of communication.

The value of β^* for $\lambda = \lambda^*$ should be above the real values if the hypothesis of minimizing $\Omega(\lambda)$ holds. All the values of β reported in Section 1 are consistent with that hypothesis when looking at Figure 3, acknowledging that the values of n and m are not known.

5 Conclusion

The fact that $P(k)$ must be a probability function imposes that $\beta > 1$ when $m \to \infty$. From the analytical point of view, varying β beyond $\beta > 1$ with $m \to \infty$ implies two additional major events: finite mean ($\langle k \rangle_P$) when $\beta > 2$ and finite variance ($\langle k^2 \rangle_P - \langle k \rangle_P^2$) when $\beta > 3$ [16]. Have the previous major events something to do with the variations of the real exponent? There are two problems. First, it is not clear why we should take $m \to \infty$. The repertoire of stimuli in humans seems finite and it is no clear why it should be large in the abnormal cases considered here. Second, the smallest and the largest real value of β ($\beta = 1.6$ and $\beta = 2.42$, respectively) lay approximately in the middle of the major events above. Therefore, the milestones above do not seem to be useful for understanding the variation of β. In contrast, considering that words have meaning, that is that signals are connected to stimuli, has the key.

Besides the previous events, there seems to be no a priori obvious reason for the particular interval of variation of β. Here we have shown that information theory may help to find the answer. To sum up, minimizing $H(S)$ could be the reason for the lower barrier in β (Fig. 4 whereas a sudden transition to $\beta \to \infty$ beyond which language is not possible, i.e. too expensive, could be the reason for the upper barrier (Fig. 5). The upper barrier can be crossed if the cost of communication is neglected, which may not be affordable by the human brain. Language may tend to maximize the information transfer but self-regulate in order to avoid paying the maximum cost, or inversely, may self-regulate in order to minimize the cost but keeping a high information transfer. The two types of schizophrenia would show what happens when that regulation fails. We support the view of language as a complex system regulating itself in order to fulfill different constraints [50,51].

If words frequencies had nothing to do with meanings, there would be more freedom for the range of real exponents. In contrast, exponents are constrained to a particular domain. Word meanings and specially the stimuli from which meaning emerges are crucial. Assuming $p(s_i) \sim \mu_i$, despite of its extreme simplicity, has a remarkable predictive power. The large amount of models for Zipf's law not assuming that words have meaning should be reconsidered [4,5,8,25–27,29,32–37,39,40]. Those models do not seem to be capable of explaining why increases in the value of β could turn in to increases in the communicative efficiency of the system (taking Figs. 1 and 3 together), something that is evident when going normal adults with $\beta = 2$ to nouns in single author texts with $\beta = 2.15 - 2.32$. In contrast, the assumption that words frequency is correlated with word number of stimuli does it naturally (recall Figs. 1 and 3). The positive correlation between communicative efficiency and β is not easy to determine, because precise information theory measures, as far as we know, have not been used for the atypical systems considered here. Those models may not be capable of explaining either the natural bounds we find in the variation of real communication systems. At present, only two models take into account that words have meanings [15,38]. Our work supports the idea that word meaning and balancing the communicative efficiency and the cost of communication are crucial for understanding word frequency distributions.

Our findings indicate that it is possible to get information about the balance between communicative efficiency and cost of communication from the word frequency distribution in a special way. The higher the exponent β, the higher the weight of the communicative efficiency, but taking into account that $\beta \to \infty$ does not imply that $\lambda = 1$, since divergence is sudden beyond a threshold value of λ.

We are grateful to Claudio Castellano, Toni Hernández and Vito Servedio for helpful comments. This work was supported by the FET Open Project COSIN, IST-2001-33555 and partially supported by the Grup de Recerca en Informàtica Biomèdica.

References

1. G.K. Zipf, *Human behaviour and the principle of least effort. An introduction to human ecology,* 1st edn. (Hafner reprint, New York, 1972) (Cambridge, MA: Addison-Wesley, 1949)
2. G.K. Zipf, *The psycho-biology of language* (Houghton Mifflin, Boston, 1935)
3. R.J. Chitashvili, R.H. Baayen, in *Quantitative Text Analysis*, edited by G. Altmann, L. Hřebíček (Wissenschaftlicher Verlag Trier, Trier, 1993), pp. 54–135
4. J. Tuldava, J. Quantitative Linguistics **3**(1), 38 (1996)
5. V.K. Balasubrahmanyan, S. Naranan, J. Quantitative Linguistics **3**(3), 177 (1996)
6. R. Ferrer i Cancho, submitted to the J. Quantitative Linguistics (2002)
7. R. Ferrer i Cancho, R.V. Solé, J. Quantitative Linguistics **8**(3), 165 (2001)

8. M.A. Montemurro, Physica A **300**, 567 (2001)
9. M.A. Montemurro, D. Zanette, Glottometrics **4**, 87 (2002)
10. R.G. Piotrowski, V.E. Pashkovskii, V.R. Piotrowski, Automatic Documentation and Mathematical Linguistics **28**(5), 28 (1995), first published in Nauchno-Tekhnicheskaya Informatisiya, Seriya 2, Vol. 28, No. 11. pp. 21–25, 1994
11. X. Piotrowska, W. Pashkovska, R. Piotrowski, to appear (2003)
12. L. Brilluen, *Science and Theory of Information* (Russian translation) (Gos. Izd-vo Fiz,-Mat. Lit-ry, Moscow, 1960)
13. G.K. Zipf, Science **96**, 344 (1942)
14. A.N. Kolguškin, *Linguistic and engineering studies in automatic language translation of scientific Russian into English. Phase II* (University of Washington Press, Seattle, 1970)
15. R. Ferrer i Cancho, Physica A **345**, 275 (2004), doi:10.1016/j.physa.2004.06.158
16. G.A. Miller, N. Chomsky, in *Handbook of Mathematical Psychology*, edited by R.D. Luce, R. Bush, E. Galanter (Wiley, New York, 1963), Vol. 2
17. L. Reder, J.R. Anderson, R.A. Bjork, J. Experimental Psychology **102**, 648 (1974)
18. R. Köhler, *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik* (Brockmeyer, Bochum, 1986)
19. C.D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, 1999), Chap. Introduction
20. S. Wolfram, *A new kind of science* (Wolfram Media, Champaign, 2002)
21. M.A. Nowak, J.B. Plotkin, V.A. Jansen, Nature **404**, 495 (2000)
22. M.A. Nowak, J. Theor. Biol. **204**, 179 (2000), doi:10.1006/jtbi.2000.1085
23. M.A. Nowak, Phil. Trans. R. Soc. Lond. B **355**, 1615 (2000)
24. W. Li, IEEE T. Inform. Theory **38**(6), 1842 (1992)
25. G.A. Miller, Am. J. Psychol. **70**, 311 (1957)
26. B. Mandelbrot, in *Readings in Mathematical Social Sciences*, edited by P.F. Lazarsfield, N.W. Henry (MIT Press, Cambridge, 1966), pp. 151–168
27. J.S. Nicolis, *Chaos and information processing* (World Scientific, Singapore, 1991)
28. R. Suzuki, P.L. Tyack, J. Buck, Anim. Behav. **69**, 9 (2005)
29. H.A. Simon, Biometrika **42**, 425 (1955)
30. D.H. Zanette, S.C. Manrubia, Physica A **295**(1-2), 1 (2001)
31. A. Rapoport, Quantitative Linguistics **16**, 1 (1982)
32. B. Mandelbrot, in *Communication theory*, edited by W. Jackson (Butterworths, London, 1953), p. 486
33. A.A. Tsonis, C. Schultz, P.A. Tsonis, Complexity **3**(5), 12 (1997)
34. I. Kanter, D.A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995)
35. S. Naranan, V.K. Balasubrahmanyan, Current Science **63**, 261 (1992)
36. S. Naranan, V.K. Balasubrahmanyan, Current Science **63**, 297 (1992)
37. S. Naranan, V.K. Balasubrahmanyan, J. Scientific Industrial Res. **52**, 728 (1993)
38. R. Ferrer i Cancho, R.V. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2003)
39. P. Harremoës, F. Topsøe, Entropy **3**, 227 (2001)
40. P. Harremoës, F. Topsøe, in *Proceedings of the International Symposium on Information Theory*, Lausanne, Switzerland (2002), p. 207
41. P. Allegrini, P. Gricolini, L. Palatella, Chaos, Solitons and Fractals **20**, 95 (2004)
42. P. Allegrini, P. Gricolini, L. Palatella, World Scientific (2003), submitted
43. A.G. Bashkirov, A.V. Vityazev, Physica A **277**, 136 (2000)
44. A.G. Bashkirov, cond-mat/0211685 (2003)
45. F. Pulvermüller, *The Neuroscience of Language. On Brain Circuits of Words and Serial Order* (Cambridge University Press, Cambridge, 2003)
46. R. Ferrer i Cancho, submitted to Phys. Rev. E (2004)
47. R. Ferrer i Cancho, F. Reina, J. Quantitative Linguistics **9**, 35
48. C.E. Shannon, Bell Systems Techn. J. **27**, 379 (1948)
49. T.J. Crow, British J. Psychiatry **173**, 303 (1998)
50. R. Köhler, Theor. Linguist. **14**(2-3), 241 (1987)
51. W. Wildgen, Recherches semiotiques - Semiotic Inquiry **14**, 53 (1989)