Eur. Phys. J. B **47**, 449–457 (2005)
DOI: 10.1140/epjb/e2005-00340-y

THE EUROPEAN PHYSICAL JOURNAL B

Zipf's law from a communicative phase transition

R. Ferrer i Cancho[a]

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

Received 14 April 2005 / Received in final form 28 July 2005
Published online 28 October 2005 –

Abstract. Here we present a new model for Zipf's law in human word frequencies. The model defines the goal and the cost of communication using information theory. The model shows a continuous phase transition from a no communication to a perfect communication phase. Scaling consistent with Zipf's law is found in the boundary between phases. The exponents are consistent with minimizing the entropy of words. The model differs from a previous model [Ferrer i Cancho, Solé, Proc. Natl. Acad. Sci. USA **100**, 788–791 (2003)] in two aspects. First, it assumes that the probability of experiencing a certain stimulus is controlled by the internal structure of the communication system rather than by the probability of experiencing it in the 'outside' world, which makes it specially suitable for the speech of schizophrenics. Second, the exponent α predicted for the frequency versus rank distribution is in a range where $\alpha > 1$, which may explain that of some schizophrenics and some children, with $\alpha = 1.5-1.6$. Among the many models for Zipf's law, none explains Zipf's law for that particular range of exponents. In particular, two simplistic models fail to explain that particular range of exponents: intermittent silence and Simon's model. We support that Zipf's law in a communication system may maximize the information transfer under constraints.

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

1 Introduction

Human word frequencies are known to obey a universal regularity. If $P(i)$ is the frequency of the ith most frequent word in a text, then it follows that

$$P(i) \sim i^{-\alpha}, \tag{1}$$

where we have typically $\alpha \approx 1$ [1,2]. Equation (1) defines the so-called Zipf's law. Equation (1) is the frequency versus rank representation of Zipf's law.

Although $\alpha \approx 1$ is usually found in word frequencies, significant deviations have been reported:

- $\alpha < 1$ in fragmented discourse schizophrenia. The speech is characterized by multiple topics and the absence of consistent subject. The lexicon of a text of that kind may be varied and chaotic [3,4]. $\alpha \in [0.7, 0.9]$ is found. That type of schizophrenia seems to be found in early stages of the disease but not all early stages should follow that pattern.
- $\alpha > 1$ in advanced forms of schizophrenia [1,3,4]. 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. $\alpha = 1.5$ is reported in [3,4].
- $\alpha > 1$ in young children [3,5,6]. $\alpha = 1.6$ is reported in [5]. Older children conform to the typical $\alpha \approx 1$ [7].
- $\alpha = 1.4$ in military combat texts [4,8].

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

Two trivial explanations have been proposed for Zipf's law: intermittent silence [9–14] and Simon's model [15]. Because of their over simplifications with regard to real words, those models are often considered null-hypothesis rather than models in the strict sense.

Intermittent silence consists of random sequences of letters (or phonemes) interrupted by blank spaces (or silences). The belief that Zipf's law in human words can be explained by a such a trivial process is widespread in science [14,16–20] although intermittent silence texts and real texts differ greatly [21,22]. The simplest intermittent silence model obeys [14,20]

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

where L is the number of letters (or phonemes) of the alphabet. Equation (2) comes from assuming that silence and each letter have probability $1/(L+1)$. It follows from equation (2) that $\alpha > 1$. The simplest intermittent silence cannot explain the cases where $\alpha < 1$ and cannot easily reproduce the frequency distribution of the cases where $\alpha > 1$. If $\alpha > 1.58$ then no simple intermittent silence model can account for the exponent (Fig. 1) which may exclude young children. If $1.26 \leq \alpha \leq 1.58$, then $L \leq 3$, which is inconsistent with the unaltered number of letters (or phonemes) of schizophrenics. Besides, schizophrenic patients with $\alpha < 1$ are also excluded because equation (2) gives $\alpha > 1$ for finite L. The exponent of military combat texts cannot be explained either because it lays between

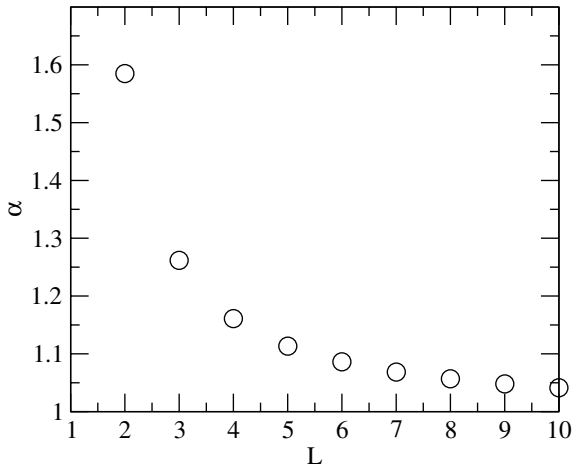

Fig. 1. α, the exponent of Zipf's law, versus L, the number of letters, of an intermittent silence model.

the exponent predicted by $L = 2$ ($\alpha = 1.58$) and $L = 3$ ($\alpha = 1.26$). A slightly more complicated intermittent silence covers continuously the values of α from 1 to ∞. If the probability of silence is σ and the probability of each letter is $1/L$ we have [10]

$$\alpha = 1 - \frac{\log(1-\sigma)}{\log L}. \tag{3}$$

Equation (2) is recovered when $\sigma = 1/(L+1)$. Assuming that L is constant (i.e. the number of letters or phonemes is unaltered in military combat texts, schizophrenics and may be also children), equation (3) predicts that the larger the value of α the larger the value of σ. In turn, the larger the value of σ, the shorter the length of words. A significant word length decrease in military combat texts and schizophrenics is hard to justify. Word length reduction is not known to be among the linguistic manifestation of schizophrenia [23]. Ultimately, the major problem intermittent silence has as a model of word frequencies is a radically unrealistic design. Intermittent silence assumes that words are created from scratch by combining characters on the fly. In contrast, real words are selected from a mental lexicon, i.e. a set of preconstructed words or base word forms [24,25]. While the mental lexicon is essentially finite, intermittent silence does not tentatively bound the number of words it can generate.

Simon's model is based on generating a text by choosing at random words that have been previously used. That model can only explain $\alpha < 1$, so both young children and some schizophrenics are excluded. Simon's original model is a birth process. Extending Simon's model so that it becomes a birth and death process [26] cannot explain exponents $\alpha > 1$ either. In sum, Simon's model or intermittent silence cannot easily explain atypical exponents. The problems are: inconsistent predictions and assumptions or parameters values that are hard to justify. The model that will be introduced here does not have such kind of problems.

The present article is devoted to show that Zipf's law with the particular range of $\alpha > 1$ found in schizophrenics and children can be explained by a non-trivial process, namely, maximizing the communicative efficiency of a system under constraints. The model presented shows Zipf's law in the vicinities of a phase transition. Phase transitions of decoding algorithms in a noisy channel have received attention in the physics literature [27–29]. The model presented here shows a phase transition between no communication and perfect communication in a noiseless channel as in [30]. We will support that Zipf's law may be a scaling law appearing in the vicinities of a phase transition [30] in different circumstances.

2 The model

We assume we have a general communication system that is defined by 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 equivalent to words and stimuli are the basic ingredients of word meaning. For instance, the word 'dog' is associated to visual stimuli (e.g. the shape of a dog), auditive stimuli (e.g. barking), ... All these stimuli are elicited by the word 'dog' [31]. Our stimuli are sometimes called objects or events in the origins of language literature [19,32]. We assume that signals link to stimuli and that connections are defined by an $n \times m$ binary matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if s_i and r_j are linked and $a_{ij} = 0$ otherwise.

We assume that the goal of human language is that of any communication system, i.e. maximizing $I(S,R)$, the Shannon's information transfer between the set of signals S and the set of stimuli R [33]. As in [30], we assume that communication has a cost of signal use that is defined by $H(S)$, the entropy associated to signals. It is known in psycholinguistics that the lower the frequency of a word, the lower its availability (the so-called word frequency effect [34]). The availability affects both the speaker when it has to find a signal for a particular stimulus and the hearer, who has to find the intended stimulus by a signal. The higher the availability the lower the cost. So the worst case situation is given by all words being equally likely ($p(s_i) = 1/n$ for each i). In that case, we have maximum $H(S)$. The minimum cost situation is when all signals have probability zero except one. In that case, $H(S) = 0$. $H(S)$ is a measure of the cost of the communication. We define the energy function Ω_0 that any communication system must minimize as in [30] as

$$\Omega_0(\lambda) = -\lambda I(S,R) + (1-\lambda)H(S), \tag{4}$$

where $0 \leq \lambda \leq 1$. λ is a parameter controlling the balance between the goals of communication (maximizing $I(S,R)$) and minimizing the cost of communication (minimizing $H(S)$). Communicative efficiency is totally favoured when $\lambda = 1$ whereas saving cost is totally favoured when $\lambda = 0$.

The information transfer between S and R can be defined in two equivalent ways [35]:

$$I(S,R) = H(R) - H(R|S) \tag{5}$$

or

$$I(S,R) = H(S) - H(S|R), \quad (6)$$

where $H(R)$ is the entropy of stimuli, $H(R|S)$ is the average entropy associated the interpretation of signals and $H(S|R)$ is the average entropy associated to choosing a certain signal for a certain stimulus. Knowing equation (5), equation (4) becomes

$$\Omega_0(\lambda) = -\lambda H(R) + \lambda H(R|S) + (1-\lambda)H(S). \quad (7)$$

Minimizing $\Omega_0(\lambda)$ when $H(R)$ is constant as in [30], keeping λ fixed, is equivalent to minimizing

$$\Omega_1(\lambda) = \lambda H(R|S) + (1-\lambda)H(S), \quad (8)$$

the main energy function used in [30]. Here we will not have constant $H(R)$.

Alternatively, we may write equation (4) as

$$\Omega_0(\lambda) = (1-2\lambda)H(S) + \lambda H(S|R) \quad (9)$$

using equation (6). Whereas $H(S|R)$ is always minimized, $H(S)$ is minimized if $1-2\lambda > 0$ and maximized if $1-2\lambda < 0$. Thus, the solution of $1-2\lambda = 0$ gives $\lambda = 1/2$ as the point where a radical change in the behavior of Ω_0 takes place.

Here we define the joint probability of s_i and r_j as in [30] as

$$p(s_i, r_j) = \frac{a_{ij} p(r_j)}{\omega_j}, \quad (10)$$

where $p(r_j)$ is the probability of the jth stimulus and

$$\omega_j = \sum_{k=1}^{n} a_{kj} \quad (11)$$

is the number of links of that stimulus. We define $\mu_i = \sum_{j=1}^{m} a_{ij}$ as the number of links of the ith signal and

$$M = \sum_{i=1}^{n} \mu_i = \sum_{j=1}^{m} \omega_i \quad (12)$$

as the total amount of links.

The main difference with [30] is that we assume

$$p(r_i) = \frac{\omega_i}{M}, \quad (13)$$

hence $H(R)$ is not constant here. Replacing equation (13) in equation (10) we obtain

$$p(s_i, r_j) = \frac{a_{ij}}{M}. \quad (14)$$

Equation (14) is used in [36]. Replacing equation (14) into $p(s_i) = \sum_{j=1}^{m} p(s_i, r_j)$ we get

$$p(s_i) = \frac{\mu_i}{M}. \quad (15)$$

Assuming $p(r_j) \sim \omega_j$ has the virtue of leading to simple probability definitions (e.g. $p(s_i) \sim \mu_i$) and also allowing one to explain the interval of variation of α in human language [37]. Interestingly, equation (14) allows stimuli with no links (provided $M > 0$) whereas equation (10) with stimulus probabilities that are independent of A does not. Here there is no freedom to determine $\{p(r_1), ..., p(r_j), ..., p(r_m)\}$ (shortly $\{p(r_j)\}$) a priori as in [30]. Here, $p(r_j)$ is a function of the matrix A. To understand the differences between the model in [30] and the present model we need to define more precisely what we mean by $p(r_j)$. $p(r_j)$ is not the probability that a stimulus of type r_j happens but the probability that a stimulus of type r_j is perceived. As an example of a reason against the former definition, let us consider the huge amount of stimuli that a normal human cannot perceive unless he uses special instruments (e.g. ultrasounds, cosmic radiation, the composition of atomic nuclei, etc.). Besides, one expects that the frequency of a word is positively correlated with the frequency of perceiving each of its associated stimuli. Equation (10) is a simple way of defining that correlation. The model in [30] assumes that $\{p(r_j)\}$ is constant. Tentatively, it seems that the probability of a certain stimulus is determined by the 'outside' world. Indirectly, that means that the frequency of a signal is strongly influenced by the frequency of its associated stimuli which is in turn externally determined. For instance, the word 'dog' is more likely to be used than the word 'aardvark' because, roughly speaking, aardvarks, edentate mammals that are common in Southern Africa, have a much more restricted habitat than dogs. Even if the word 'dog' and the word 'aardvark' where connected in the same way in A, the frequency of the signal 'dog' should be higher than the frequency of the signal 'aardvark'.

That strong link between the 'outside' world and signal use is challenged by displaced reference, the ability to talk about something that is distant in time or space [38,39]. Displaced reference is an important feature of human language but it is not uniquely human since bees have it [40]. Because of displaced reference, we can talk of 'dogs' even when the stimuli elicited by dogs are not coming from the 'outside' world. The probability of experiencing the stimuli associated to the word 'dog' is mostly the probability of talking or thinking about dogs. The latter probability is related to that of experiencing dogs in the 'outside' world but the relationship is not necessarily direct. If the flow between 'inside' and 'outside' world is direct, one would expect that internal stimulus probabilities mirror external ones. We do not know about the extent to which a direct flow exists in normal adults but there are special cases where a purely direct flow seems clearly less likely. What could happen when the boundary between the self and the 'outside' world is lost, as in schizophrenia [41,42]? Various core aspects of the disease such as false believes, hallucinations [41] and various cognitive impairments, including attention problems [43], indicate that the relationship with the 'outside' world is enormously complicated.

When we assume equation (13), we are assuming that the probability of using a certain signal comes from 'inside' but in a special way, i.e. from the internal organization of the communication system. That is what we hypothesize specially for schizophrenia. That also could be happening

in young children whose relationship with the 'outside' world is under construction.

$H(S)$ and $H(S|R)$ in equation (9) can be calculated with the only assumption of equation (14). From the one hand, the standard information theory definition [35]

$$H(S) = -\sum_{i=1}^{n} p(s_i) \log p(s_i) \tag{16}$$

can be developed using equation (15), which leads to

$$H(S) = \log M - \frac{1}{M} \sum_{i=1}^{n} \mu_i \log \mu_i. \tag{17}$$

From the other hand, the standard information theory definition [35]

$$H(S|R) = \sum_{j=1}^{m} p(r_j) H(S|r_j) \tag{18}$$

can be developed using equation (13) and knowing that $H(R|s_i) = \log \mu_i$ [36] so $H(S|r_j) = \log \omega_j$. Thus, we obtain

$$H(S|R) = \frac{1}{M} \sum_{j=1}^{m} \omega_j \log \omega_j. \tag{19}$$

As for $H(S|R)$, its standard information theory definition, [35]

$$H(R|S) = \sum_{i=1}^{n} p(s_i) H(R|s_i), \tag{20}$$

can be developed using equation (15) and knowing that $H(R|s_i) = \log \mu_i$ [36]. Since $H(R|S)$ is the same as $H(S|R)$ changing ω_j by μ_i, we get

$$H(R|S) = \frac{1}{M} \sum_{i=1}^{n} \mu_i \log \mu_i. \tag{21}$$

Replacing the previous equation in equation (17), we get

$$H(S) = \log M - H(R|S). \tag{22}$$

3 Results

Here we minimize $\Omega(\lambda)$ keeping n and m constant for different values of λ. A Monte Carlo technique at zero temperature for minimizing $\Omega(\lambda)$ is used as in [30]. The algorithm is based on generating random changes in A and choosing any change decreasing $\Omega(\lambda)$. In [30], the number of changes in A follows a binomial distribution. The particular issue here is simplifying the minimization algorithm by making constant the number of changes in A. Here we take the number of changes to be exactly two which is similar to the two expected changes in [30]. The number of changes should not be too large since otherwise the minimization algorithm cannot converge to any solution. If the number of changes is one (the smallest possible), then the changes can only be of two types (a) add a new link or (b) remove an existing link. Changes keeping constant the amount of links are not allowed, which is too restrictive. Choosing two simultaneous changes gives (a) add a new link and remove an existing link (b) add two new links (c) remove two existing links. Changes keeping the amount of links constant are allowed. Taking too many changes may jeopardize the convergence to a minimum of $\Omega(\lambda)$, so we take only two fixed changes for simplicity.

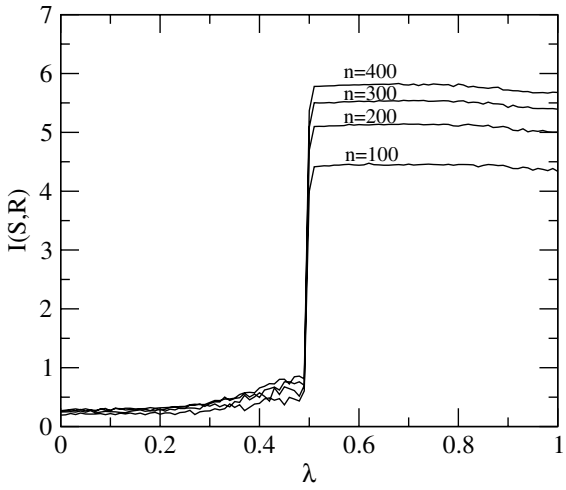

Fig. 2. $I(S,R)$ versus λ for systems of different sizes: $n = m = 100$, $n = m = 200$, $n = m = 300$ and $n = m = 400$. $I(S,R)$ is the information transfer between the set of signals (S) and the set of stimuli (R). λ is the parameter regulating the balance between maximizing the information transfer and saving the cost of signal use. Averages over 30 realizations are shown. Natural logarithms are used for $I(S,R)$.

Since the information transfer can be equivalently defined as equations (6) and (5), it follows that the maximum value of $I(S,R)$ is given by the maximum value of $H(S)$, $\log n$, and the maximum value of $H(R)$, $\log m$. Since equations (6) and (5) are equivalent, it follows that $I(S,R) \leq \log \min(n,m)$. Therefore, if a configuration reaches $I(S,R) \approx \log \min(n,m)$ we will say that the information transfer is maximum and the system is an (almost) perfect communicator.

Figure 2 shows $I(S,R)$ versus λ for systems minimizing $\Omega(\lambda)$ and having different sizes. A sudden jump from the minimum information transfer to the maximum information transfer is found for $\lambda = \lambda^*$. The type of abrupt change suggests a phase transition from no communication to perfect communication. λ^* is defined as the approximate point where scaling is found.

Figure 3 shows that $P(i)$ is changing from a fast decaying function of i to a rather flat curve in a very narrow interval of λ. In between, a distribution consistent with equation (1) is found, supporting that a continuous phase transition [44] is taking place for $\lambda \approx \lambda^* = 1/2 - \epsilon$ where ϵ is a small positive number.

Since a Zipf's law-like distribution is found for $\lambda^* < 1/2$, that is, when $H(S)$ is minimized, we wonder if the

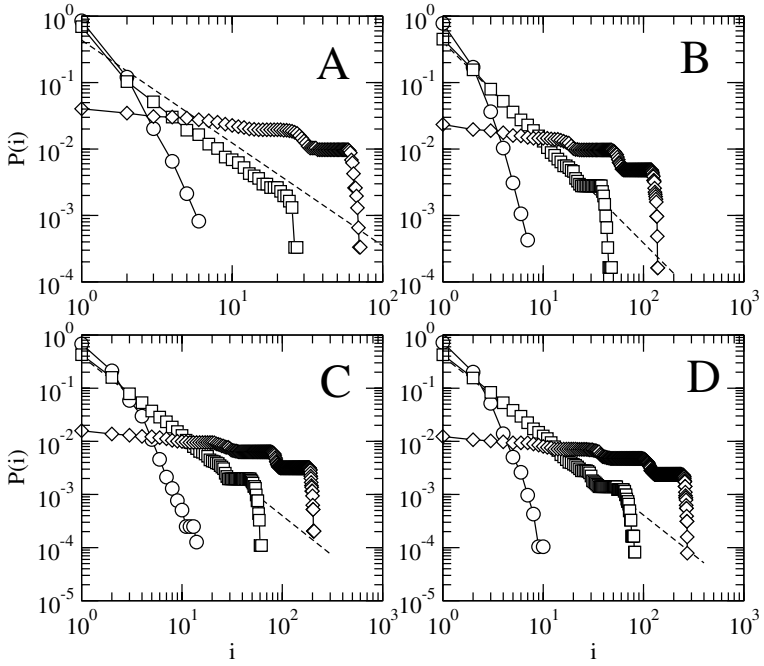

Fig. 3. $P(i)$, the probability of the ith most frequent signal, obtained from minimum energy configurations for systems of different sizes: $n = m = 100$ (A), $n = m = 200$ (B), $n = m = 300$ (B) and $n = m = 400$ (D). Four series are shown in each plot: $\lambda = 0.49$ (circles), $\lambda = \lambda^*$ (squares), $\lambda = 1/2$ (diamonds) and the ideal curve for α^*, the value of α minimizing $H(S)$ when $\mu_i \sim i^{-\alpha}$ (dashed line). μ_i is the number of links of the ith most connected signal. Averages of $P(i)$ over 30 realizations are shown. When $\lambda = \lambda^*$ we have $\alpha^* = 1.54$ for $n = m = 100$, $\alpha^* = 1.51$ for $n = m = 200$, $\alpha^* = 1.5$ for $n = m = 300$ and $\alpha^* = 1.49$ for $n = m = 400$. λ^* is determined approximately. We have chosen $\lambda^* = 0.4986$ for $n = m = 100$, $\lambda^* = 0.4987$ for $n = m = 200$, $\lambda^* = 0.4987$ for $n = m = 300$ and $\lambda^* = 0.4986$ for $n = m = 400$.

value of α we obtain can be explained by that minimization and assuming only

$$\mu_i = ci^{-\alpha}, \tag{23}$$

where c is a constant, so that equation (1) follows using equation (15). We can calculate $H(S)$ using equations (17) and (23). Equation (17) depends on c, n and α. Finding α^*, the value of α minimizing $H(S)$, when c and n are constant implies a search on a two-dimensional landscape. We would like to be able to reduce the number of dimensions of that landscape. We can use a different definition of Zipf's law, that is

$$P(f) \sim f^{-\beta}, \tag{24}$$

where $P(f)$ is the proportion of signals whose probability is f. $P(f)$ is the so-called frequency spectrum representation of Zipf's law [45]. It is known that equations (1) and (24) are equivalent with [46]

$$\beta = 1/\alpha + 1 \tag{25}$$

and $\beta > 1$. Assuming equation (15) we may write equation (24) as

$$P(k) \sim k^{-\beta}, \tag{26}$$

where $P(k)$ is the proportion of signals with k links. We define $\langle ... \rangle$ as the expectation operator over

$$\{P(1), ..., P(k), ..., P(m)\}. \tag{27}$$

Using $M = n \langle k \rangle$ and $P(k)$, we may write equation (17) as

$$H(S) = \log(n \langle k \rangle) - \frac{1}{n \langle k \rangle} \sum_{k=1}^{m} nP(k)k \log k. \tag{28}$$

After some algebra we get

$$H(S) = \log(n \langle k \rangle) - \frac{\langle k \log k \rangle}{\langle k \rangle}, \tag{29}$$

which depends on n,m and β. Minimizing equation (29) with n and m constant is equivalent to minimizing

$$H'(S) = \log \langle k \rangle - \frac{\langle k \log k \rangle}{\langle k \rangle}, \tag{30}$$

where $\langle k \rangle$ and $\langle k \log k \rangle$ depend only on m and β. We may find β^*, the value of β minimizing $H'(S)$ for different values of m. Once β^* is obtained, we can use equation (25) to get α^*. Figure 4A shows $H(S)$ versus α with $n = m$ for the values of m used in Figure 3. We have $\alpha^* = 1.54$ for $n = m = 100$, $\alpha^* = 1.51$ for $n = m = 200$, $\alpha^* = 1.50$ for $n = m = 300$ and $\alpha^* = 1.49$ for $n = m = 400$. Values of α^* are close to the exponents obtained when minimizing $\Omega(\lambda)$ for $\lambda = \lambda^*$ (Fig. 3). The previous results suggest that the scaling found is due to a continuous phase transition where the exponent α is such that minimizes $H(S)$. Figure 4B shows that α^* decays with m very slowly for sufficiently large m, suggesting that if $1 < \alpha < 2$ is found

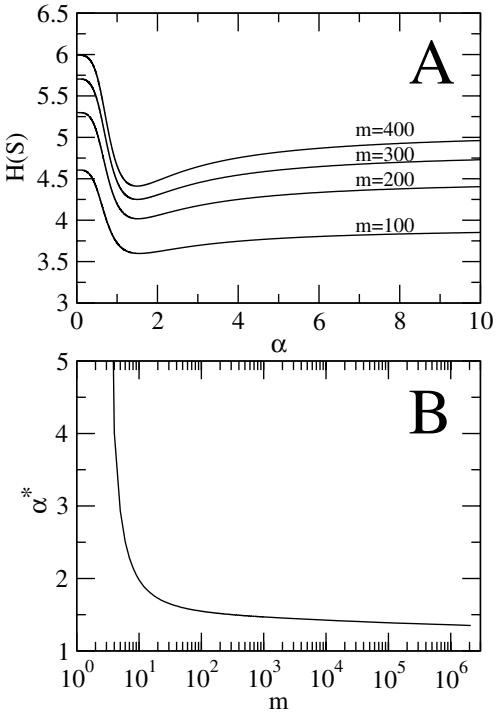

Fig. 4. (A) $H(S)$, the signal entropy, versus α, the exponent of the frequency versus rank representation of Zipf's law. Series with $n = m$ and different values of m are shown. Natural logarithms are used for $H(S)$. (B) α^*, the value of α minimizing $H(S)$ versus m, the number of stimuli.

in a communication systems that could indicate $H(S)$ is being minimized. Furthermore, if $\alpha \leq 1$ then $H(S)$ is not being minimized.

4 Discussion

We have seen that minimizing equation (4) can lead to Zipf's law at the vicinities of a phase transition with a particular range of exponents. Nonetheless, we have not discussed if alternative equations may or may not exhibit that behavior. Tentatively, minimizing $\Omega_0(\lambda)$ (Eq. (4)) is not equivalent to minimizing $\Omega_1(\lambda)$ (Eq. (8)) unless $H(R)$ is constant. $H(R)$ is constant in the model in [30] but is not here, so studying $\Omega_1(\lambda)$ is needed. Equation (8) could be rewritten as

$$\Omega_1(\lambda) = (1 - 2\lambda)H(S) + \lambda \log M, \tag{31}$$

knowing $H(R|S) = \log M - H(S)$ (Eq. (22)). Equation (31) is the same as equation (9) changing $\log M$ by $H(S|R)$. A radical change in behavior is expected for $\lambda = 1/2$.

We could replace $H(S)$ by a simpler measure of cost that would be proportional to the effective lexicon size, i.e. the number of signals with at least on connection. Thus, we define

$$J(X) = |\{x|x \in X \text{ and } d(x) > 0\}|, \tag{32}$$

where $|...|$ is the cardinality operator and $d(x)$ is the degree or number of connections of x. $J(X)$ could take S or R as parameters. We define another possible measure of the cost of the lexicon S as

$$L_n(S) = \frac{J(S) \log n}{n}. \tag{33}$$

$L_n(S)$ is a rescaled measure of the effective lexicon size. $L_n(S)$ is designed to vary in the same range as $H(S)$, that is, from 0 to $\log n$. Therefore, we replace $H(S)$ by $L_n(S)$ in equation (4) and obtain

$$\Omega_2(\lambda) = -\lambda I(S, R) + (1 - \lambda)L_n(S). \tag{34}$$

Finally, replacing $H(S)$ by $L_n(S)$ and $-I(S, R)$ by $H(R|S)$ in equation (4) we obtain

$$\Omega_3(\lambda) = \lambda H(R|S) + (1 - \lambda)L_n(S). \tag{35}$$

Using $\Omega_1(\lambda)$ we found a sharp transition for $\lambda = 1/2$ similar to that of $\Omega_0(\lambda)$ (Eq. (4)). No scaling was found suggesting that the transition is discontinuous. The absence of scaling using $\Omega_1(\lambda)$ suggests that $\log M$ is too simple for showing Zipf's law. No sudden transition and no scaling was found for either $\Omega_2(\lambda)$ and $\Omega_3(\lambda)$. The previous results suggest that $L_n(S)$ is too simple for showing a transition. The same was found using $L_n(S)$ instead of $H(S)$ in [30], where $\{p(r_j)\}$ is constant.

We end our seek of alternative energy functions by considering another energy function

$$\Omega_4(\Lambda) = -I(S, R) + \Lambda H(S), \tag{36}$$

where Λ is constant controlling the weight of $H(S)$. Although it may not be obvious at first glance, Ω_4 is obtained from Ω_0 through a simple algebraic transformation: $\Omega_4(\Lambda) = \Omega_0(\lambda)/\lambda$ and $\Lambda = (1 - \lambda)/\lambda$. The outcome of minimizing $\Omega_4(\Lambda)$ is the same as that of minimizing $\Omega_0(\lambda)$ with $\Lambda = (1 - \lambda)/\lambda$. Whereas $\Omega_0(\lambda)$ has a sharp transition for $\lambda = 1/2$, $\Omega_4(\Lambda)$ has it for $\Lambda = 1$.

Among the large amount explanations for Zipf's law in human words: tautologies [1,47], birth process [15], word length minimization [11,13,48], intermittent silence [10,11,14,48], differential equations for the word frequency distribution [45,49,50], various types of random Markov processes [13,51], optimization of information theory measures [2,36,52–54], communicative phase transitions [30], entropy discontinuities [55,56], random walks in complex networks [57,58] and general models for scaling [59,60], none, as far as we know, explains Zipf's law for the particular range of exponents found in children and schizophrenics. The model presented here does it (Figs. 3, 4). It has two different phases: a no communication phase (where $I(S, R) \approx 0$) for $\lambda < 1/2 - \epsilon$ and a perfect communication phase (where $I(S, R)$ approaches the maximum value) for $\lambda > 1/2$. Just at the transition point, $\lambda = \lambda^* = 1/2 - \epsilon$ a configuration (not a phase) consistent with atypical human language is found. For values of λ sufficiently far from the transition point, $P(i)$ is characterized by different exponents: $\alpha \to \infty$ in the no

communication phase, α finite but significantly different from $\alpha = 0$ in the transition point and $\alpha \approx 0$ (i.e. all signals are equally likely) at the perfect communication phase. Interestingly, $\alpha = 0$ (or near), the exponent maximizing $I(S,R)$, is never found in human language. We may think that Zipf's law in word frequencies has nothing to do with communication, specially when some trivial mechanisms that have nothing to do with communication can explain some intervals of α [9–11,13–15]. The key to understanding that human language or any other communication system showing Zipf's law actually maximizes $I(S,R)$ is that the cost of communication prevents language from achieving the ideal $\alpha = 0$.

The model in [30] also shows Zipf's law on the edge of a communicative transition, but $\alpha \approx 1$ is obtained instead of $\alpha \approx 1.5$ as here. The differences in the exponent suggests that using stimulus probabilities independent of A instead of stimulus probabilities depending on A could make a substantial difference in the behavior of a communication system. The model presented here has some advantages with regard to the model presented in [30]:

- Algebraic simplicity, e.g. $H(R|s_i) = \log \mu_i$ and $H(S|r_j) = \log \omega_j$. Thus, straightforward arguments about the expected value of α at the transition can be made (Fig. 4).
- Stimulus probabilities are not a parameter. We have $p(r_j) \sim \omega_j$ (Eq. (13)).
- More freedom: unlinked stimuli are allowed.
- Faster calculations: the static calculation of Ω takes $\Theta(\max_i\{\mu_i\}) + \max_j\{\omega_j\})$ time (whereas $\Theta(\sum_{k=1}^{n} \mu_k)$ time is needed in [30]).

The same numerical results we have found here could be obtained by the analytical model in [36]. There, equation (13) is also assumed. The model introduced here obtains Zipf's law when minimizing $\Omega_0(\lambda)$ with $\lambda = \lambda^*$ whereas the model in [36] does so when the entropy of $\{P(k)\}$ is maximized and $\langle \log k \rangle$ is constrained. $\langle \log k \rangle$ comes from the fact that if $\mu_i = k$ then $H(R|s_i) = k$ [36]. $H(R|s_i)$ is the uncertainty associated to decoding s_i, in other words, a measure of the cost of decoding s_i. While the model presented here gives a particular range of exponents that depends on m, the model in [36] leaves the exact value of the exponent to additional constraints. Since the model here and the model in [36] share the same probability definitions, the latter can explain explain Zipf's law with the same dependence between the exponents and m if $H(S)$ minimization is assumed once the system organizes according to equation (26).

Exponents very close to $\alpha = 1.5$ are found in the model (Figs. 3, 4) as well as in the speech of young children [3,5] and schizophrenic patients [3,4]. The coincidence supports that assuming $p(r_j) \sim \omega_j$ is suitable for the case where the relationship between the 'inside' and 'outside' world is different than that of normal adults. Thus, word frequencies of children and schizophrenic with $\alpha > 1$ may be tuned to maximize communicative efficiency at a critical point where word entropy is minimized. Signal entropy minimization makes sense in young children whose less developed brain imposes minimizing the cost of communication, as well as in schizophrenics, where the illness seems to affect the distinction between the speaker and the hearer [61].

We have seen that our model could explain $\alpha \approx 1.5$ found in children and one type of schizophrenics but what can we say about the other type, where $\alpha < 1$? The normal flow between the 'inside' and the 'outside' world is generally altered in schizophrenia so it makes sense to think that the assumption $p(r_j) \sim \omega_j$ (Eq. (13)) keeps having sense for patients with $\alpha < 1$. The difference could be that those patients may not be subject to a critical balance between maximizing the information transfer and the cost of communication described by equation (4). The model in [36] could explain them since it leaves the exact value of the exponent to later constraints. More precisely, the problem of schizophrenics with $\alpha < 1$ could be that they are paying an excessive cost for communicating. Knowing that cost of communication (in terms of $H(S)$) grows as α decreases when $\alpha < \alpha^*$ (Fig. 4A), schizophrenics with $\alpha < 1$ could be paying a higher cost than those with $\alpha > 1$. Additionally, the fact that the cost of decoding a signal decreases with α (because it increases with β [36]; recall Eq. (25)) suggests that the individual words of schizophrenics with $\alpha < 1$ are more accurate than those of schizophrenics and children with $\alpha > 1$. In sum, schizophrenics and children with $\alpha > 1$ seem to have the most refined cost-benefit tuning allowed by the assumption $p(r_j) \sim \omega_j$ whereas schizophrenics with $\alpha < 1$ seem far from that.

Normal adult speech is in between no attachment and total attachment to the 'here and now'. Detachment and attachment may not have the same weight. It is reasonable to think that the amount of detachment from the present increases the chance that the normal interaction between the 'inside' and 'outside' world is interrupted. The repeated use of a word too independently from the present could eventually dissociate the internal probability of the associated stimuli from their 'outside' probabilities. It is reasonable to think that this could be happening somehow in normal adult speakers. The extreme situation in normal adults could be military combat texts, where α deviates clearly from the typical value found in normal adults. We do not mean that military texts and schizophrenics are equivalent but they may share the fact the communication system is apparently controlling the probability of each stimulus. Differences between military combat texts and schizophrenics do exist. For instance, if we assume that $H(S)$ is minimized, then the expected value of m for each case differ in various orders of magnitude ($m \approx 300$ is expected of $\alpha = 1.5$ while $m \approx 30\,000$ is expected for $\alpha = 1.4$). The fact that schizophrenic and military combat texts are presumably extremely goal oriented (i.e. one topic) cannot explain alone what may be happening in schizophrenia and military combat texts. Normal adults can focus on a particular topic but maintain the typical exponent. The difference between normal adults and those special cases may not be a matter of quantity. A very large value of m, a way of neglecting topic constraints, does not give $\alpha^* \approx 1$, what we typically find in normal adults. If we take $m = 3.35 \times 10^7$, $\alpha^* = 1.32$

would be found, which is still far from the typical $\alpha \approx 1$. α^* grows slowly with m (Fig. 4). Besides, the altered link with the 'outside' world needs to be considered as one of the possible explanations for some writers. For instance, $\alpha = 1.6$ is found in the complete Shakespeare works [2]. That exponent cannot easily be attributed to the fact that Shakespeare is doing fiction work. David Copperfield by Charles Dickens gives $\alpha = 1.20$ [62], Don Quixote by Miguel de Cervantes gives $\alpha = 1.05$ [62], the Aeneid by Virgil gives $\alpha = 0.68$ [62] and the Ulysses by James Joyce gives $\alpha = 1.05$ [2]. Shakespeare's low exponent suggests that the author may have unconsciously let the communication system to take control on the frequency of stimuli.

A combination of broken and unbroken flows between the 'inside' and the 'outside' world makes sense in normal adults. In that case, the model in [30] and the model introduced here represent the two extremes. One where stimulus probabilities are separated from the structure of the communication system. There the probability of a stimulus is dictated by the probability of perceiving the stimulus in the 'outside' world or by the probability of thinking about it (perceiving it in the 'inside' world). That would be the model in [30]. Another one where the communication system has taken the whole control. That would be the present model or the model in [36]. The advantage of the present model over the model in [36] is that it uses the same energy function that can explain the typical exponent of Zipf's law in world languages. The suitability of the present model and the model in [36] for explaining the atypical exponents in schizophrenia and children suggests the internal organization of the communication system has a lot of influence in perceiving stimulus in those cases. Further work is needed to understand the differences between the the present model and the model in [30].

Some caution should be taken interpreting the results of the present article. First, $p(r_j) \sim \omega_j$ may not be the only way of accounting for atypical exponents. It could be that the model in [30] accounts for the atypical exponents with a small modification or a particular set of parameters. Second, $p(r_j) \sim \omega_j$ makes special sense in schizophrenics and young children but the possibility that it is also suitable for apparently normal speakers, maybe only in special cases, cannot be denied. If the assumption $p(r_j) \sim \omega_j$ is valid for communication systems, it is reasonable to think that it is so at least in schizophrenics and young children. Future work should focus on the nature of $p(r_j)$ in communication systems.

We thank Toni Hernández, Brita Elvevåg and Claudio Castellano for helpful comments. The author also thanks two anonymous referees for sharp criticisms. This work was funded by the FET Open Project COSIN, IST-2001-33555 and the ECAgents project, funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-1940. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of the data appearing in this publication.

References

1. G.K. Zipf, *Human behaviour and the principle of least effort. An introduction to human ecology* (Hafner reprint, New York, 1972), 1st edn. (Cambridge, MA, Addison-Wesley, 1949)
2. V.K. Balasubrahmanyan S. Naranan, J. Quantitative Linguistics **3**(3), 177 (1996)
3. R.G. Piotrowski, V.E. Pashkovskii, V.R. Piotrowski, Automatic Documentation and Mathematical Linguistics **28**(5), 28 (1995), first published in Naučno-Techničeskaja Informatisija, Serija 2 **28**, No. 11, 21 (1994)
4. X. Piotrowska, W. Pashkovska, R. Piotrowski, to appear (2003)
5. L. Brilluen, *Science and theory of information, Russian translation* (Gos. Izd-vo Fiz,-Mat. Lit-ry, Moscow, 1960)
6. B. McCowan, L.R. Doyle, S.F. Hanser, J. Comparative Psychology **116**, 166 (2002)
7. G.K. Zipf, Science **96**, 344 (1942)
8. A.N. Kolguškin, *Linguistic and engineering studies in automatic language translation of scientific Russian into English. Phase II* (University of Washington Press, Seattle, 1970)
9. R. Suzuki, P.L. Tyack, J. Buck, Anim. Behav. (2003), accepted
10. G.A. Miller, Am. J. Psychol. **70**, 311 (1957)
11. B. Mandelbrot, in *Readings in mathematical social sciences*, edited by P.F. Lazarsfield, N.W. Henry (MIT Press, Cambridge, 1966), pp. 151–168
12. 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
13. J.S. Nicolis, *Chaos and information processing* (World Scientific, Singapore, 1991)
14. W. Li, IEEE T. Inform. Theory **38**(6), 1842 (November 1992)
15. H.A. Simon, Biometrika **42**, 425 (1955)
16. S. Wolfram, *A new kind of science* (Wolfram Media, Champaign, 2002)
17. M.A. Nowak, J.B. Plotkin, V.A. Jansen, Nature **404**, 495 (2000)
18. M.A. Nowak, J. Theor. Biol. **204**, 179 (2000)
19. M.A. Nowak, Phil. Trans. R. Soc. Lond. B **355**, 1615 (2000)
20. R. Suzuki, P.L. Tyack, J. Buck, Anim. Behav. **69**, 9 (2005)
21. A. Cohen, R.N. Mantegna, S. Havlin, Fractals **5**(1), 95 (1997)
22. R. Ferrer i Cancho, R.V. Solé, Adv. Complex Syst. **5**, 1 (2002)
23. L.E. DeLisi, Schizophrenia Bulletin **27**(3) (2001)
24. D.W. Carroll, *Psychology of language* (Brooks/Cole Publishing Company, Pacific Grove, California, 1994)
25. A. Akmajian, R.A. Demers, A.K. Farmer, R.M. Harnish, *Linguistics. An Introduction to Language and Communication* (MIT Press, 1995)
26. S. Manrubia, D. Zanette, J. Theor. Biol. **216**, 461 (2002)
27. S. Franz, M. Leone, A. Montanari, F. Ricci-Tersenghi, Phys. Rev. E **66**, 046120 (2002)
28. A. Montanari, N. Sourlas, Eur. Phys. J. B **18**, 107 (2000)
29. A. Montanari, Eur. Phys. J. B **18**, 121 (2000)
30. R. Ferrer i Cancho, R.V. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2003)

31. F. Pulvermüller, *The neuroscience of language. On brain circuits of words and serial order* (Cambridge University Press, Cambridge, 2003)
32. R. Ferrer i Cancho, O. Riordan, B. Bollobás, Proc. R. Soc. Lond. Series B **272**, 561 (2005)
33. C.E. Shannon, Bell Systems Technical J. **27**, 379 (1948)
34. *Handbook of Psycholinguistics*, edited by M.A. Gernsbacher (Academic Press, San Diego, 1994)
35. R.B. Ash, *Information Theory* (John Wiley & Sons, New York, 1965)
36. R. Ferrer i Cancho, Physica A **345**, 275 (2004)
37. R. Ferrer i Cancho, Eur. Phys. J. B **44**, 249 (2005)
38. N. Chomsky, Paper presented at the Univ. de Brasilia, Nov. 26 (1996)
39. C.F. Hockett, *A course in modern linguistics* (McMillan, New York, 1958)
40. K. von Frisch, Scientific American **207**, 79 (1962)
41. K.T. Mueser, S.R. McGurk, The Lancet **363**, 2063 (2004)
42. T.J. Crow, Schizophrenia Research **28**, 127 (1997)
43. B. Elvevåg, T.E. Goldberg, Critical Reviews in Neurobiology **14**, 1 (2000)
44. J. Binney, N. Dowrick, A. Fisher, M. Newman, *The theory of critical phenomena. An introduction to the renormalization group* (Oxford University Press, New York, 1992)
45. J. Tuldava, J. Quantitative Linguistics **3**(1), 38 (1996)
46. R.J. Chitashvili, R.H. Baayen, in *Quantitative Text Analysis*, edited by G. Altmann, L. Hřebíček (Wissenschaftlicher Verlag Trier, Trier, 1993), pp. 54–135
47. A. Rapoport, Quantitative Linguistics **16**, 1 (1982)
48. B. Mandelbrot, in *Communication theory*, edited by W. Jackson (Butterworths, London, 1953), p. 486
49. M.A. Montemurro, Physica A **300**, 567 (2001)
50. A.A. Tsonis, C. Schultz, P.A. Tsonis, Complexity **3**(5), 12 (1997)
51. I. Kanter, D.A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995)
52. S. Naranan, V.K. Balasubrahmanyan, Current Science **63**, 261 (1992)
53. S. Naranan, V.K. Balasubrahmanyan, Current Science **63**, 297 (1992)
54. S. Naranan, V.K. Balasubrahmanyan, J. Scientific and Industrial Research **52**, 728 (1993)
55. P. Harremoës, F. Topsøe, Entropy **3**, 227 (2001)
56. P. Harremoës, F. Topsøe, in *Proceedings of the International Symposium on Information Theory*, Lausanne, Switzerland (2002), p. 207
57. P. Allegrini, P. Gricolini, L. Palatella, Chaos, solitons and fractals **20**, 95 (2004)
58. P. Allegrini, P. Gricolini, L. Palatella (World Scientific, 2003), submitted
59. A.G. Bashkirov, A.V. Vityazev, Physica A **277**, 136 (2000)
60. A.G. Bashkirov, e-print: `cond-mat/0211685` (2003)
61. T.J. Crow, British J. Psychiatry **173**, 303 (1998)
62. M.A. Montemurro, D. Zanette, Glottometrics **4**, 87 (2002)