Decoding least effort and scaling in signal frequency distributions

Ramon Ferrer i Cancho [a,b]

[a] *INFM udR Roma1, Dip. di Fisica.*
Universita "La Sapienza",
Piazzale A. Moro 5
00185 Roma, Italy.
Phone: 00 39 06 4455705
Fax: 00 39 06 4452045

[b] *Complex Systems Lab,*
Universitat Pompeu Fabra (UPF),
Dr. Aiguader 80, 08003 Barcelona, Spain

PACS 89.75.-k,01.30.-y

Abstract

Here, assuming a general communication model where objects map to signals, a power function for the distribution of signal frequencies is derived. The model relies on the satisfaction of the receiver (hearer) communicative needs when the entropy of the number of objects per signal is maximized. Evidence of power distributions in a linguistic context (some of them with exponents clearly different from the typical $\beta \approx 2$ of Zipf's law) is reviewed an expanded. We support the view that Zipf's law reflects some sort of optimization but following a novel realistic approach where signals (e.g. words) are used according to the objects (e.g. meanings) they are linked to. Our results strongly suggest that many systems in nature use non-trivial strategies for easing the interpretation of a signal. Interestingly, minimizing just the number of interpretations of signals does not lead to scaling.

Key words: Zipf's law, scaling, human language, animal communication

We assume a general communication framework where signals are mapped to the objects they refer to [1,2]. For vervet monkeys, we have alarms calls as signals and predators as objects [3]. For human language, we have words as signals and meanings as objects, acknowledging that meaning is a complex matter to define [4] and we humans make use of symbolic reference and not indexical reference as many animals seem to do [5]. For Unix computer commands, we have commands and their options as signals and the computer

operations they imply as objects [6]. For the immune system, we have reactivity patterns as signals and antigens as objects [7,8]. We assume communication takes place between an ideal sender (speaker) and an ideal receiver (hearer). The task of the sender is to code an object using a signal that the receiver has to decode [9].

Word frequencies are usually modeled by a power function

$$p_f \sim f^{-\beta} \tag{1}$$

where p_f is the proportion of words whose frequency is f in a given sample text and $\beta > 0$. The regularity is called Zipf's law honoring G. K. Zipf, the linguist who made it popular [10]. We typically have $\beta \approx 2$ [11,12] but slight variations around β have been recorded [11]. There are some interesting clear deviations:

(1) Schizophrenia with $1 < \beta < 2$ [10].
(2) Variations in the exponent when focusing on certain types of words (Fig. 1). We find $\beta = 3.35$ for English nouns (Fig. 1 B) [13] whereas we find $\beta = 1.94$ for English verbs (Fig. 1 A)
(3) The peripheral lexicon [12,14]. Studies on multiauthor collections of texts show two domains in p_f. One domain with with $\beta \approx 2$ for the most frequent words and another domain with $\beta \approx 3/2$ for the less frequent words. The two regimes are said to divide words into a core and peripheral lexicon, respectively. We assume we focus on the core lexicon in the present paper.
(4) Shakespearean ouvres with $\beta = 1.6$ [11]. This a rather controversial situation because Shakespearean works are likely to be a case of multiauthorship [15] and thus show the shape of a peripheral lexicon. What follows must be cautiously interpreted for this case.

Besides human language, scaling consistent with Zipf's law is found in the frequency of immune reactivity patterns [7,8] and the computer commands issued by experienced Unix users with $\beta = 2.24$ [6] (Fig. 1 B, [16]).

We are aimed at answering the following questions:

(1) Is there any general principle allowing to explain whatever form scaling in signal frequency distributions?
(2) Is such a principle totally different from any explanation for the typical $\beta \approx 2$?
(3) How does information transfer depends on β?

Many explanations have been proposed for scaling in word frequencies [17–29]. Most of such models assume a certain optimization or stability principle. Given

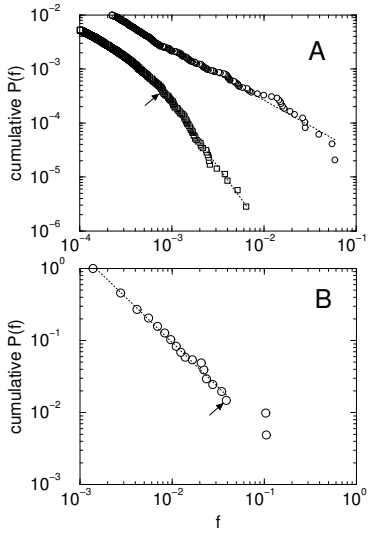

Fig. 1. $P(f)$ the probability a signal has normalized frequency f in cumulative form. Power approximations are shown for every series (dotted lines). Arrows indicate the point considered as the end of the straight line for calculating the exponents β. A. p_f for English verbs with $\beta = 1.94 \pm 0.003$(circles) and English nouns with $\beta = 3.35 \pm 0.02$(squares). The core lexicon starts slightly before $f \approx 10^{-3}$ B. Unix computer commands issued by an experienced user $\beta = 2.24 \pm 0.028$.

the large amount of models, Zipf's law models require a discussion not only about the suitable models but also about distinguishing the causes of Zipf's law (the true model(s)) from its consequences (the side-effect models). This is not the aim of the present paper. Nonetheless, all the explanations for Zipf's law (except [17]) forget a fundamental reason for which words are used: words are used according to their meaning. Real sentences are not a collection of words entirely chosen at random as many models intend [19,18,20,22]. Following the approach in [17], we assume that words are chosen according to their meaning and that the frequency of a word is a function of the objects eliciting it.

Recently, it has been shown that G. K. Zipf's proposal of a principle of least effort for the hearer and the speaker can explain $\beta \approx 2$ [17]. In a few words, G. K. Zipf proposed that Zipf's law results from a trade-off between hearer and speaker needs. In G. K. Zipf's rough intuition, the sender prefers a few words for all meanings (unification) and the hearer needs every meaning has a different word (diversification). The higher the degree of satisfaction of the needs of one of them, the less its effort. The model in [17] uses a parameter λ for minimizing $\Omega = \lambda E_D + (1-\lambda)E_C$, a linear combination of E_D, the coding effort (the effort for the hearer/receiver) and E_C, the coding effort (the effort for the speaker/sender), with $0 \leq \lambda \leq 1$. Sender and receiver needs are totally satisfied when $\lambda = 0$ and $\lambda = 1$, respectively. A phase transition separates a no communication phase (sender's full satisfaction) and a perfect communication phase (receiver's full satisfaction). Scaling consistent with Zipf's law with $\beta \approx 2$ is found at some intermediate value of $\lambda = \lambda^*$. The model shows a

continuous phase transition [30] at the point where sender and receiver needs are at the maximum tension. We will refer to this model as the *dual least effort satisfaction model*. Here we show that scaling in word frequencies can be explained only complying with receiver needs under a convenient maximization principle. We will refer to this model as the *decoding least effort model*.

We assume we have a set of signals $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of objects of reference $R = \{r_1, .., r_j, ..., r_m\}$. We define a matrix of signal-object associations $A = \{a_{ij}\}$ $(1 \leq i \leq n \ , \ 1 \leq j \leq m)$ where $a_{ij} = 1$ if the i-th signal and the j-th object are connected and $a_{ij} = 0$ otherwise. Here, we define the joint probability of the i-th signal and the j-th object as

$$p(s_i, r_j) = \frac{a_{ij}}{\sum_{k=1}^{n} \mu_k}$$

where μ_i, the number of objects linked to the i-th signal, is defined as

$$\mu_i = \sum_{j=1}^{m} a_{ij}. \tag{2}$$

Knowing the frequency of the i-th signal is

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{3}$$

we obtain

$$p(s_i) = \frac{\mu_i}{\sum_{k=1}^{n} \mu_k}.$$

The probability of understanding r_j when s_i is received is

$$p(r_j|s_i) = \frac{p(s_i, r_j)}{p(s_i)}$$

so we have

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

The probability definitions used here are simpler than in [17].

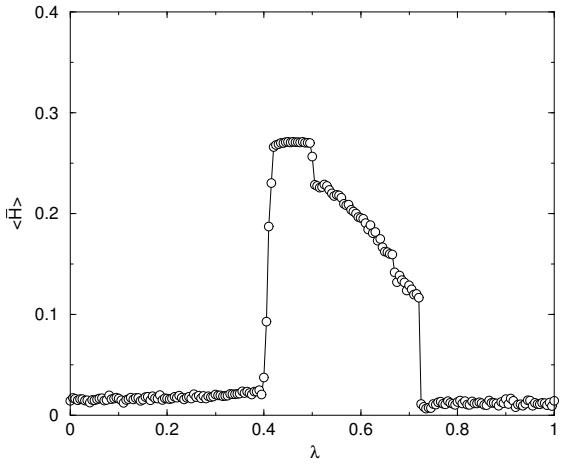

Fig. 2. $< \bar{\mathcal{H}} >$ the mean normalized entropy of the number of objects per signal (solid line) versus λ, where $\bar{\mathcal{H}} =< \mathcal{H} > / \log m$. For $\lambda \approx 0.41$ as sharp transition takes place and scaling is found in the dual least effort model (n=m=150).

We define $\mathcal{H}$, the entropy of the number of objects per signal, as

$$\mathcal{H} = -\sum_{k=1}^{m} p_k \log p_k$$

where

$$p_k = \frac{|\{i|\mu_i = k \; and \; 1 \leq i \leq n\}|}{n}.$$

The maximization principle we will use for E_D comes from the observation that $\mathcal{H}$ is maximal at the point where scaling is found in [17] (Fig. 2). Thus, we can obtain $\{p_k\} = (p_1, ..., p_k, ..., p_m)$ using the maximum entropy principle [31,32]. We define $\Phi(k)$ as the decoding effort (effort for the receiver) implied once a signal linked to k objects has been issued. We seek $\{p_k\}$ maximizing the *a priori* uncertainty $\mathcal{H}$ under the decoding effort we define here as

$$E_D = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mu_i) \tag{5}$$

and the normalization constraint

$$\sum_{k=1}^{m} p_k = 1. \tag{6}$$

Rewriting Eq. 5 as

$$E_D = \sum_{k=1}^{m} p_k \Phi(k) \tag{7}$$

we end up with the functional

$$\Omega = \mathcal{H} + \alpha \sum_{k=1}^{m} p_k + \beta \sum_{k=1}^{m} p_k \Phi(k).$$

The distribution $\{p_k\}$ maximizing $\mathcal{H}$ will be deduced from the condition $\partial\Omega/\partial p_k = 0$ which leads to different distributions depending on Φ. Once s_i has been issued, the receiver must avoid interpreting an object that was not intended by the sender. The simplest way of satisfying the receiver needs is just minimizing μ_i, which leads to $\Phi(k) = k$ when $\mu_i = k$. A more sophisticated strategy consists of minimizing $H(R|s_i)$, the entropy of objects when s_i is given, defined as

$$H(R|s_i) = -\sum_{j=1}^{n} p(r_j|s_i) \log p(r_j|s_i). \tag{8}$$

$H(R|s_i)$ measures the uncertainty associated to the interpretation of s_i.

Replacing Eq. 4 into Eq. 8 we get

$$H(R|s_i) = -\sum_{j=1}^{m} \frac{a_{ij}}{\mu_i} \log \frac{a_{ij}}{\mu_i}$$

which gives $H(R|s_i) = \log \mu_i$. According to $H(R|s_i)$, if the i-th word has $\mu_i = k$ objects, then it implies an effort $\Phi(k) = \log k$.

For $\Phi(k) = k/m$ and large m, $\frac{\partial\Omega}{\partial p_k} = 0$ leads to [33]

$$p_k \sim e^{-k/<k>} \tag{9}$$

where c is a normalization term. For $\Phi(k) = \log k$, we obtain [31]

$$p_k \sim k^{\beta'}. \tag{10}$$

$\beta' < 0$ is satisfied provided that [31]

$$\frac{\sum_{k=1}^{m} k^{\beta'} \log k}{\sum_{k=1}^{m} k^{\beta'}} < \frac{1}{m} \sum_{k=1}^{m} \log k.$$

Zipf's law [10] can be straightforwardly obtained from Eq. 10 with $\beta' = -2$. If f is the frequency of a signal and p_f is the probability of f, Eq. 3 can be written as

$$f = \frac{k}{n \sum_{i=1}^{m} p_k k} = \frac{k}{n \langle k \rangle}$$

If $P(k = K)$ is the probability the random variable k (the number of objects per signal) is K then using $p_f = P(k = fn \langle k \rangle)$ with Eq. 10 we get

$$p_f \sim f^{\beta'}.$$

Using the same argument on Eq. 9 we obtain

$$p_f \sim e^{-nf}.$$

We have seen that explaining a wide range of exponents for the scaling in word frequencies is a relaxation of a more restrictive principle, minimizing both the coding and decoding effort. The dual least effort satisfaction model predicts that all signals will tend to have the same frequency if only receiver needs are satisfied. The decoding least effort presented here, with scaling in word frequencies, does not contradict the dual least effort model. The decoding least effort model assumes what is a side-effect close to the phase transition in the dual least effort model, i.e. maximizing $\mathcal{H}$.

We have seen that the decoding least effort model with $\Phi(k) = \log k$ predicts without specifying the value of β. The dual least effort model shows scaling consistent with Zipf's law for $\lambda \approx \lambda^*$ [17]. When $\lambda < \lambda^*$, word frequencies obey

$$P(i) \approx \begin{cases} 1 \; if \; i = 1 \\ 0 \; otherwise \end{cases} \tag{11}$$

where $P(i)$ is the frequency of the i-th most frequent word. Eq. 11 can be rewritten as $P(i) \sim i^{-\alpha}$ with $\alpha \to \infty$. When $\lambda > \lambda^*$, word frequencies obey $p(i) \sim 1/n$ which can be rewritten as $P(i) \sim i^{-\alpha}$ with $\alpha = 0$. Knowing (see for instance [34,35,12])

$$\beta = \frac{1}{\alpha} + 1, \tag{12}$$

we can argue that Eq. 1 is always present in the dual least effort model, not only for the transition but also at the two phases. The typical Zipf's law in

human language is a particular case of scaling with non-extreme exponents, since $P(i) \sim i^{-\alpha}$ is only monotonically decreasing (and thus $P(i)$ can be defined as the frequency of the i-th most frequent word) only when $\alpha \in [0, \infty)$.

Now, we will find a simple relationship between E_D and β. c, the normalization term of Eq. 10, can be approximated solving

$$c \int_1^m k^{-\beta} dk = 1 \tag{13}$$

which leads to

$$c \approx \frac{1-\beta}{m^{1-\beta}-1} \tag{14}$$

provided $\beta \neq 1$. β can be approximately determined substituting Eq. 10 into the definition of E_D of Eq. 7 as follows

$$E_D = \int_1^m ck^{-\beta} \log k dk \tag{15}$$

Solving the integral in the right side of the previous equation with $\beta \neq 1$ we get

$$E_D = c\frac{1}{1-\beta}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right) + \frac{1}{1-\beta}\right] \tag{16}$$

which we rewrite as

$$E_D = \frac{1}{m^{1-\beta}-1}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right) + \frac{1}{1-\beta}\right] \tag{17}$$

using Eq. 14. Notice that the previous equation is undetermined for $\beta = 1$ or $m = 1$. If $m \to \infty$ and $\beta > 1$ we have

$$\beta = \frac{1}{E_D} + 1 \tag{18}$$

It follows from Eq. 12 and Eq. 18 that $E_D = \alpha$. Since $E_D \geq 0$ (when $\beta > 1$), then solving

$$\frac{dE_D}{d\beta} = -\frac{2}{(\beta-1)^2} = 0$$

gives a global minimum of E_D for $\beta \to \infty$.

Knowing that $\alpha = 0$ and therefore $\beta \to \infty$ minimize not only E_D but also maximize the potential information transfer [36,37], we may ask why human language has chosen $\alpha \approx 1$ and therefore $\beta \approx 2$ as its typical exponents. Is the answer that human language is more a system of thought and mental representation than a communication system as some researchers have proposed [38–40]? Probably the answer is that the pressure for maximizing information transfer, minimizing the decoding effort in human language has to satisfy conflicting goals. The dual least effort model tells us that adding coding least effort is a suitable answer for $\beta \approx 2$. Other exponents require putting into consideration other constraints. Nouns, with $\beta \approx 3.35$ are closer to the theoretical maximum information limit ($\beta \to \infty$), suggesting they have violated the balance or or maximum tension between coding and decoding needs in other to achieve higher information transfer, that is, lower decoding effort. Eq. 18 suggests that nouns have lower decoding effort than the typical E_D given by Zipf's law with $\beta \approx 2$. Similarly, schizophrenics speech with $1 < \beta \leq 2$ suggest they are not taking into account the effort for the hearer their exponents imply high values of E_D. This is consistent with the suspect that schizophrenic speakers tend to lump together too many meanings in one form of expression. Schizophrenics overload word meanings [10].

Therefore, exponents are indicators of E_D and have to do with information transfer. To make it more explicit, Eq. 1 and Eq. 18 give

$$p_f \sim f^{-\frac{1}{E_D}-1}.$$

It should be understood from the present work that $\beta < \infty$ does not imply that scaling in word frequency has nothing to do with effective communication although different mechanisms can lead to Zipf' law [37]. Eq. 18 bridges the gap between power word frequency distributions and communicative efficiency. There are many possible ways of minimizing the decoding effort, but probably only one where hearer and speaker needs are at the maximum tension, i.e. $\beta \approx 2$.

Our work puts a step forward to understand complex and simpler communications systems. The former making use of $\Phi(k) = \log k$ and the latter $\Phi(k) = k$. Scaling in different contexts [10,7,8,6] suggests that many systems in nature make use of non-trivial mechanisms for reducing the uncertainty associated to the codes they generate. Minimizing $\Phi(k) = k$ helps to decrease the uncertainty associated to the interpretation of a signal but does not lead scaling.

Acknowledgements

We thank Ryuji Suzuki for helpful discussions. This work was supported by the *Institució Catalana de Recerca i Estudis Avançats (ICREA)*, the *Grup de Recerca en Informàtica Biomèdica (GRIB)*, and grants of the Generalitat de Catalunya (FI/2000-00393).

References

[1] M. A. Nowak, D. C. Krakauer, The evolution of language, Proc. Natl. Acad. Sci. USA 96 (1999) 8028–8033.

[2] M. A. Nowak, J. B. Plotkin, D. C. Krakauer, The evolutionary language game, J. theor. Biol. 200 (1999) 147–162.

[3] R. M. Seyfarth, D. Cheney, P. Marler, Vervet monkey alarm calls: semantic communication in a free-ranging primate, Anim. Behav. 28 (1980) 1070–194.

[4] Y. Ravin, C. Leacock (Eds.), Polysemy. Theoretical and computational approaches, Oxford University Press, New York, 2000.

[5] M. D. Hauser, The evolution of communication, MIT Press, Cambridge, MA, 1996.

[6] S. R. Ellis, R. J. Hitchcock, The emergence of Zipf's law: spontaneous encoding by users of a command language, IEEE Trans. Syst. Man Cyber. 16 (3) (1986) 423–427.

[7] J. D. Burgos, Fractal representation of the immune b cell repertoire, BioSystems 39 (1996) 19–24.

[8] J. D. Burgos, P. Moreno-Tovar, Zipf-scaling behavior in the immune system, BioSystems 39 (1996) 227–232.

[9] R. B. Ash, Information Theory, John Wiley & Sons, New York, 1965.

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

[11] V. K. Balasubrahmanyan, S. Naranan, Quantitative linguistics and complex system studies, J. Quantitative Linguistics 3 (3) (1996) 177–228.

[12] R. Ferrer i Cancho, R. V. Solé, Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited, J. Quantitative Linguistics 8 (3) (2001) 165–173.

[13] Frequencies
obtained from A. Kilgarriff's word-frequency list of the British National corpus (http://www.itri.brighton.ac.uk/∼Adam.Kilgarriff/bnc-readme.html).

[14] R. Ferrer i Cancho, Core and peripheral lexicon through word length optimization, Submitted to the Journal of Quantitative Linguistics .

[15] J. F. Michell, Who wrote Shakespeare?, Thames & Hudson, Slovenia, 1999.

[16] Statistics perfomed on the *bash* history file of an anonymous experienced user at the Complex Systems Lab.

[17] R. Ferrer i Cancho, R. V. Solé, Least effort and the origins of scaling in human language, Proc. Natl. Acad. Sci. USA 100 (2003) 788–791.

[18] B. Mandelbrot, An informational theory of the statistical structure of language, in: W. Jackson (Ed.), Communication theory, Butterworths, London, 1953, p. 486.

[19] H. A. Simon, On a class of skew distribution functions, Biometrika 42 (1955) 425–440.

[20] G. A. Miller, Some effects of intermittent silence, Am. J. Psychol. 70 (1957) 311–314.

[21] J. S. Nicolis, Chaos and information processing, World Scientific, Singapore, 1991.

[22] W. Li, Random texts exhibit Zipf's-law-like word frequency distribution, IEEE T. Inform. Theory 38 (6) (1992) 1842–1845.

[23] S. Naranan, V. K. Balasubrahmanyan, Models for power law relations in linguistics and information science, J. Quantitative Linguistics 5 (1-2) (1998) 35–61.

[24] P. Harremoës, F. Topsøe, Zipf's law, hyperbolic distributions and entropy loss, in: IEEE International Symposium on Information Theory, 2002, in press.

[25] W. Li, Letters to the editor, Complexity 3 (1998) 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).

[26] M. A. Montemurro, Beyond the Zipf-Mandelbrot law in quantitative linguistics, Physica A 300 (2001) 567–578, cond-mat/0104066.

[27] L. Pietronero, E. Tosatti, V. Tosatti, A. Vespignani, Explaining the uneven distribution of number in nature: the laws of Benford and Zipf, Physica A 293 (2001) 297–304.

[28] S. Denisov, Fractal binary sequences: Tsallis thermodynamics and the Zipf's law, Phys. Lett. A 235 (1997) 447–451.

[29] A. G. Bashkirov, A. V. Vityazev, Information entropy and power-law distribution for chaotic systems, Physica A 277 (2000) 136–145.

[30] 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.

[31] J. N. Kapur, Maximum entropy models in science and engineering, Wiley, New Delhi, 1989, Ch. Maximum-entropy discrete univariate probability distributions, pp. 30–43.

[32] E. W. Montroll, M. F. Shlesinger, Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails, J. Stat. Phys. 32 (1983) 209–230.

[33] H. Haken, Synergetics an introduction: nonequilibrium phase transitions & self-Organization in physics, chemistry & biology, Springer-Verlag, New York, 1979.

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

[35] S. Naranan, V. K. Balasubrahmanyan, Information theoretic models in statistical linguistics - Part I: A model for word frequencies, Current Science 63 (1992) 261–269.

[36] T. M. Cover, J. A. Thomas, Elements of information theory, Wiley, New York, 1991.

[37] R. Suzuki, P. L. Tyack, J. Buck, The use of Zipf's law in animal communication analysis, Anim. Behav. Accepted.

[38] N. Chomsky, Aspects of the theory of syntax, MIT Press, Cambridge, MA, 1965.

[39] D. Bickerton, Language and species, Chicago University Press, 1990.

[40] R. Jackendoff, Patterns in the mind, Basic Books, 1994.