On the universality of Zipf's law for word frequencies

Ramon Ferrer i Cancho

"The occurrence of Zipf's law does not constitute evidence of some powerful and universal psychological force that shapes all human communication in a single mould" (Miller & Chomsky 1963).

1 Introduction

It is hard to imagine how the development of quantitative linguistics would have been after G.K. Zipf's untimely death without the work of G. Altmann. This article aims to honour a living giant of the Zipfian school of linguistics, and presents some findings that contradict the opening statement of Miller & Chomsky that has undermined Altmann's scientific tradition for decades. But first, what is Zipf's law for word frequencies? Word frequencies arrange themselves according to Zipf's law (Zipf 1949), that the frequency of the *i*-th most frequent word in a text obeys approximately

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

The mathematical form of equation (1) is often called a power law (Newman 2005). As far as we know, Zipf's law holds in all languages where it has been tested. Given the apparent universality of Zipf's law and also the enormous differences between all languages on Earth, it is tempting to think that its explanation has nothing to do with language.

The dominant view has been that Zipf's law originates from a trivial process (Rapoport 1982, Miller & Chomsky 1963) and this view continues to dominate the scientific culture (Suzuki et al. 2005, Wolfram 2002). Since it is known that a random sequence of letters including blanks behaving as word delimiters, reproduce Zipf's law for word frequencies (Miller 1957, Mandelbrot 1966), this argument of intermittent silence has been often used for questioning the relevance of Zipf's law. Intermittent silence has recurrently been used to argue against the relevance, meaningfulness and utility of Zipf's

law in human language and other communication systems (Rapoport 1982, Suzuki et al. 2005, Wolfram 2002). Indeed, the fact that intermittent silence reproduces Zipf's law cannot be questioned[1] but as will be discussed its suitability for real human language is questionable. Intermittent silence assumes that sequences of words are uncorrelated (i.e. a word appears independently of other words). In contrast, syntax is responsible to a great extent for the existence of correlations between words within real word sequences (Ferrer i Cancho & Elevåg 2005). Thus, it is striking that those who have largely defended syntax as the crux of human language (Hauser et al. 2002) argue that intermittent silence can explain Zipf's law in real human language. Simon's (1955) model has a similar problem because it generates uncorrelated sequences of words. Either syntax is not the crux of human language or intermittent silence is not a good model. Wisdom suggests the latter option. A further weakness of intermittent silence as an explanation is that it covers only $\alpha > 1$, while $\alpha < 1$ is often found in real language (Ferrer i Cancho & Servedio, Ferrer i Cancho 2005b). For many other inconsistent predictions made by intermittent silence see, for example (Newman 2005, Ferrer i Cancho & Elevåg 2005, Ferrer i Cancho 2005c).

Since trivial explanations for Zipf's law fail, what kind of explanation should we expect? The fact that languages on Earth exhibit enormous differences is still very important. The explanation should contain ingredients that have to do with language specifically but at the same time be shared by all world languages. First, all languages have in common the fact that they serve communication and it is hard to imagine a reliable communication system that does not maximize information transfer. Second, all languages are produced by brains that need to save energy when communicating. One possible way of defining the cost of word use is the availability of words for psychological processes such as finding the appropriate word for a certain meaning (Brown & McNeil 1966) or recognizing a word (Connine et al. 1990). It is well-known that word availability is subject to the so-called word frequency effect, that states that the higher the frequency of a word, the higher its availability (Akmajian et al. 1995). The entropy of words has been proposed as a measure of the cost of word use (Ferrer i Cancho 2005c). When all words are equally likely, all words have the lowest frequency. This is the worst case for word availability and the signal entropy is maximum. When one word is

---

1. Although there are some technical problems such as the way intermittent silence fills the frequency spectrum (Ferrer i Cancho & Solé 2002).

used for everything, this is the best case for word availability and the word entropy takes its minimum value[2]. In sum, the key ingredients we propose are maximizing the information transfer and minimizing the cost of word use. In the present article, we review a family of models where the interplay between these two leads to Zipf's law for word frequencies. We will emphasise the assumptions and implications.

2   The family of models

The family of models assumes we have a set of n words $S = \{s_1, \ldots, s_i, \ldots, s_n\}$ that communicate about a set of m meanings $R = \{r_1, \ldots, r_j, \ldots, r_m\}$. A is a binary matrix indicating which word-meaning pairs are connected. A word s_i and a meaning r_j are connected if $a_{ij} = 1$ (otherwise $a_{ij} = 0$). A defines the structure of the communication system[3].

   We define $I(S, R)$ as the information transfer between words and meanings and $H(S)$ as the entropy of words, which as mentioned above, is a measure of the cost of words use. We define Ω as the function that a communication system has to minimize. A possible definition of Ω is[4]

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

where λ is a parameter controlling the balance between maximizing the information transfer and minimizing the cost of word use. λ ranges from 0 to 1. When $\lambda = 0$, all the weight is put on minimizing the cost of word use. When $\lambda = 1$, all the weight is put on maximizing the information transfer.

---

2. The entropy of words is not only a measure of the speaker effort (as argued in Ferrer i Cancho & Solé (2003) from a narrow interpretation of G.K. Zipf's hypothesis) but also a measure of the hearer's effort. This is because the word-frequency effect not only concerns word production (Brown & McNeil 1966) but also word recognition (Connine et al. 1990).

3. We choose the term meaning because it is the easiest to understand by a general audience. The framework can be made more general by defining S as a set of signals and R as a set of states. States could be meanings, objects/events, stimuli or mental internal states. What words actually communicate about is an open question (Elman 2005). The current framework is abstract enough to allow the unsatisfied reader to replace S and R by his/her preferences.

4. Equation (2) is apparently the most general communication function that leads to Zipf's law. Other Eqs. such as $\Omega(\lambda) = \lambda H(R|S) - (1 - \lambda)H(S)$ work for the model in Ferrer i Cancho & Solé (2003) but not for that in Ferrer i Cancho (2005c).
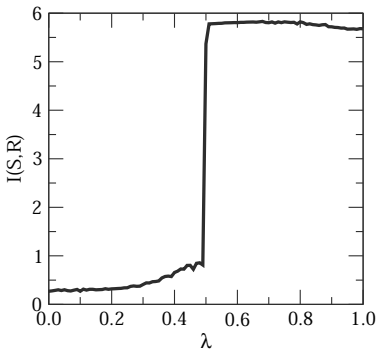
(a) $I(S,R)$, the information transfer be-
tween words and meanings, versus λ,
the parameter regulating the balance
between maximizing $I(S,R)$ and mini-
mizing the entropy of words.
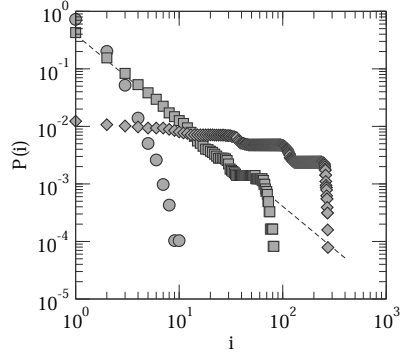
(b) $P(i)$, the probability of the i-th most
likely word in the system for $\lambda = 0.49$
(circles), $\lambda = 0.498$ (squares) and $\lambda = 0.5$ (diamonds). The dashed line con-
tains the theoretical curve for $\lambda = 0.498$. See Ferrer i Cancho (2005c)
for further details on this figure.

Figure 1: Some computational results on the model where meaning probabilities are
governed by the internal structure of the communication system. The size
of the system is $n = m = 400$ (i.e. 400 words and meanings).

In order to complete the description of the framework, we need to define
the probabilities that are used for calculating $\Omega(\lambda)$ from equation (2)[5]. We
define $p(s_i)$ and $p(r_j)$ as the probability of s_i and r_j, respectively. We de-
fine μ_i and ω_j, as the number of connections of s_i and r_j, respectively. More
precisely, we have

$$\mu_i = \sum_{j=1}^{m} a_{ij} \tag{3}$$

and

$$\omega_j = \sum_{i=1}^{n} a_{ij}. \tag{4}$$

---

5. Space precludes a full explanation of how $\Omega(\lambda)$ is calculated in depth. We just provide
the essential probabilities that allow one to calculate $\Omega(\lambda)$ using standard probability and
information theory. Further details can be found in Ferrer i Cancho (2005c) and Ferrer i
Cancho & Solé (2003).

The total amount of connections of the communication system is defined as

$$M = \sum_{i=1}^{n} \mu_i. \tag{5}$$

We define $p(s_i|r_j)$ as the probability of producing s_i for r_j.

Various recent models about Zipf's law (Ferrer i Cancho 2005a,c, Ferrer i Cancho & Solé 2003) share the fundamental assumption that

$$p(s_i|r_j) = \frac{a_{ij}}{\omega_j}. \tag{6}$$

We define $p(s_i, r_j)$ as the joint probability of s_i and r_j. Bayes theorem gives

$$p(s_i, r_j) = p(s_i|r_j)p(r_j), \tag{7}$$

which can be written as

$$p(s_i, r_j) = \frac{a_{ij}p(r_j)}{\omega_j} \tag{8}$$

using equation (6). Equation (8) is the point at which the different models diverge. $p(r_j)$ can be determined a priori (Ferrer i Cancho & Solé 2003) or from the structure of the communication system (Ferrer i Cancho 2005a,c). As for the first option, $p(r_j)$ is fixed. Ferrer i Cancho & Solé (2003) study the particular case $p(r_j) = 1/m$. As for the second option, it is assumed that $p(r_j) = \omega_j/M$, which leads to

$$p(s_i) = \frac{\mu_i}{M} \tag{9}$$

using

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{10}$$

in equation (8).

The two branches of models are very interesting from the philosopher's perspective in that one assumes that the frequency of what we talk about is dictated by the 'outside' world while the other leaves the frequency to the internal organization of the communication system itself. Tentatively, the first branch may seem more reasonable, but in fact, communication in human language is often detached from the here and now (Hockett 1958).

When speaker and hearer are near to each other, humans adults do not tend to talk about things that are near the hearer or happening right now, maybe because communicating is not very useful when the speaker and hearer are having similar sensory experiences. Therefore, the models assuming that $p(r_j)$ is not fixed suggest a possible way to study displaced reference, although it is hard to establish from the state of the art of cognitive science whether displaced speech acts are entirely controlled by the internal structure of the communication system or not.

$\Omega(\lambda)$ can be minimized using a simple Monte Carlo algorithm.[6] The general outcome of $\Omega(\lambda)$ minimization in the family of models will be illustrated using the model where meaning probabilities are not determined a priori. Figure 1 A shows that a sudden jump in information transfer takes place at a critical value of λ such that $\lambda = \lambda^* = 1/2 - \varepsilon$, where ε is a small positive value ($\varepsilon \approx 0.002$ in Figure 1). The behavior of $\Omega(\lambda)$ in the model where meaning probabilities are determined a priori is qualitatively similar[7]. The radical differences between frequency versus rank distribution of near values of λ can be seen in Figure 1a. Zipf's law is found at the sharp increase in $I(S,R)$ at $\lambda \approx 1/2$.

3   Discussion

Our model is not only interesting for philosophers but also for physicists. This is because the presence of Zipf's law near the transition point suggests that a continuous phase transition is taking place between a "no communication phase" ($I(S,R) \approx 0$ when $\lambda < \lambda^*$) and a "perfect communication phase" ($I(S,R) \approx \log min(n,m)$ when $\lambda > \lambda^*$). Phase transitions are common phenomena in nature. The transformation of boiling water into gas is maybe one of the most popular examples. The family of models visited here sheds new light on the complexity of language: language could be a system self-organizing itself between order and disorder as many other complex systems (Langton 1990, Kauffman 1993). A fully ordered configuration is one where one word is used for everything ($\lambda < \lambda^*$). A fully disordered configuration is

---

6. Details about the minimization algorithm can be found in (Ferrer i Cancho 2005c, Ferrer i Cancho & Solé 2003).

7. The are some differences. In the model reviewed here (Ferrer i Cancho 2005c): (a) the growth of $I(S,R)$ does not show an intermediate plateau near $\lambda \approx 1/2$ and (b) the transition point seems to be located closer to $\lambda = 1/2$.

one where all words are equally likely ($\lambda > \lambda^*$). Zipf's law is something in between. Indeed, our models suggest that language may operate on the edge of complete disorder: a small increase in λ in a communication system at the transition point could radically throw the system into a fully disordered domain where the maximum cost of word use is expended.

It can be argued that Zipf's law is recovered in a domain where the tendency for regularity is actually a driving force. We have seen that Zipf's law is recovered when $\lambda < 1/2$. Equation (2) minimizes $H(S)$ only when $\lambda < 1/2$[8]. If the assumptions of our models are correct, that means that human languages minimize $H(S)$, which has implications for alternative models. For instance, Mandelbrot devised an explanation for word frequencies based on maximizing $H(S)$ and constraining the mean word length (Mandelbrot 1966). However, it is hard to imagine how a real communication system whose brain maximized the entropy of words, since that would imply that the cost of word use is being maximized. Therefore, our framework questions the realism of other models and narrows down further the set of realistic explanations for Zipf's law for word frequencies.

Figure 1b shows that the effective vocabulary size (the number of words with non-zero probability) is much smaller than the potential vocabulary size. Although the potential lexicon size is $n = 400$, less than 100 words have non-zero probability at the point where Zipf's law is found. A reduced effective vocabulary size is a side-effect of the entropy minimization at the transition point. Interestingly, it has been shown that replacing $H(S)$ by the effective lexicon size (i.e. the amount of words with at least one connection) precludes the emergence of Zipf's law in the models reviewed here (Ferrer i Cancho 2005c, Ferrer i Cancho & Solé 2003). This is a key point in the understanding of the fundamental communication principles behind Zipf's law. Standard information theory (where the goal of a communication system is only maximizing $I(S,R)$) – cf. Ash (1965) – has been very successful in engineering problems but needs to be extended to apply to natural communication systems. Notice that maximizing $I(S,R)$ alone would lead to a flat probability distribution (i.e. $\alpha \approx 0$) [9].

---

8. $H(S|R)$ is the conditional entropy of words when meanings are known. Knowing that $I(S,R) = H(S) - H(S|R)$ (Ash 1965), equation (2) can be transformed into

$$\Omega(\lambda) = (1 - 2\lambda)H(S) + \lambda H(S|R). \tag{11}$$

Thus, $H(S)$ is minimized when $\lambda < 1/2$ and maximized when $\lambda > 1/2$.

9. Similar to $P(i)$ for $\lambda = 1/2$ in Figure 1b

Briefly, our models suggest that
– The entropy of words is minimized.
– Vocabulary size reduction is a side-effect of minimizing the cost of word use.
– Zipf's law for word frequencies could be the manifestation of a complex system operating between order and disorder.
– Natural communication systems require the use of extended information theory.

Contrary to Miller & Chomsky (1963), there might be in fact a single mould for all languages on Earth. The interplay between maximizing the information transfer and saving the cost of communication may constrain the possible communication systems to the subset of communication systems following Zipf's law for word frequencies. The weakness of simple explanations of Zipf's law and the family of models examined here suggest that G.K. Zipf's hypotheses about the nature of the law that bears his name (Zipf 1949) were pointing in the right direction.

References

Akmajian, Adrian; Demers, Richard A.; Farmer, Ann K.; Harnish, Robert M.
  1995   *Linguistics. An Introduction to Language and Communication.* Cambridge, Mass.: MIT Press.
Ash, Robert B.
  1965   *Information Theory.* New York: John Wiley & Sons.
Brown, Roger; McNeill, David
  1966   "The 'tip of the tongue' phenomenon". In: *Journal of Verbal Learning and Verbal Behaviour*, 5; 325–337.

Connine, Connine M.; Mullennix, John; Shernoff, Eve; Yelen, Jennifer
  1990   "Word familiarity and frequency in visual and auditory word recognition". In: *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16; 1084–1096.
Elman, Jeffrey L.
  2005   "An alternative view of the mental lexicon". In: *Trends in Cognitive Sciences*, 8; 301–306.
Ferrer i Cancho, Ramon
  2005a   "Decoding least effort and scaling in signal frequency distributions". In: *Physica A*, 345; 275–284.
  2005b   "The variation of Zipf's law in human language". In: *European Physical Journal B*, 44; 249–257.
  2005c   "Zipf's law from a communicative phase transition". In: *European Physical Journal B*, 47; 449–457.
Ferrer i Cancho, Ramon; Elvevåg, Brita
  2005   "Can intermittent silence explain Zipf's law for word frequencies?" *[Submitted]*.
Ferrer i Cancho, Ramon; Servedilo, Vito D.P.
  2005   "Can simple models explain Zipf's law for all exponents?" In: *Glottometrics*, 11; 1–8.
Ferrer i Cancho, Ramon; Solé, Ricard V.
  2002   "Zipf's law and random texts". In: *Advances in Complex Systems*, 5; 1–6.
  2003   "Least effort and the origins of scaling in human language". In: *Proceedings of the National Academy of Sciences USA*, 100; 788–791.
Hauser, Marc D.; Chomsky, Noam; Fitch, W. Tecumseh
  2002   "The faculty of language: what is it, who has it and how did it evolve?" In: *Science*, 298; 1569–1579.
Hockett, Charles F.
  1958   *A course in modern linguistics*. New York: McMillan.
Kauffman, Stuart A.
  1993   *The Origins of Order: Self-Organization*. New York: Oxford University Press.
Langton, Chris G.
  1990   "Computation at the edge of chaos: phase transitions and emergent computation". In: *Physica D*, 42; 12–37.
Mandelbrot, Benoit
  1966   "Information theory and psycholinguistics: a theory of word frequencies". In: Lazarsfield, Paul F.; Henry, Neil W. (Eds.), *Readings in mathematical social sciences*. Cambridge, Mass.: MIT Press, 151–168.

Miller, George A.
  1957   "Some effects of intermittent silence". In: *American Journal of Psy-chology*, 70; 311–314.
Miller, George A.; Chomsky, Noam
  1963   "Finitary models of language users". In: Luce, Robert D.; Bush, Robert R.; Galanter, Eugene (Eds.), *Handbook of Mathematical Psychology, vol. 2*. New York: Wiley, 419–491.
Newman, Mark E.J.
  2005   "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics*, 46; 323–351.
Rapoport, Anatol
  1982   "Zipf's law re-visited". In: *Quantitative Linguistics*, 16; 1–28.
Simon, Herbert A.
  1955   "On a class of skew distribution functions". In: *Biometrika*, 42; 425–440.
Suzuki, Ryuji; Tyack, Peter L.; Buck, John
  2005   "The use of Zipf's law in animal communication analysis". In: *Animal Behaviour*, 69; 9–17.
Wolfram, Stephen
  2002   *A new kind of science*. Champaign: Wolfram Media.
Zipf, George Kingsley
  1935   *The psycho-biology of language*. Boston: Houghton Mifflin.
  1949   *Human behaviour and the principle of least effort. An introduction to human ecology*. Cambridge, Mass.: Addison-Wesley.