Glottometrics 11, 2005, 1-8

Can simple models explain Zipf's law for all exponents?

Ramon Ferrer i Cancho[*]*, Rome*
Vito D. P. Servedio, Rome

Abstract. H. Simon proposed a simple stochastic process for explaining Zipf's law for word frequencies. Here we introduce two similar generalizations of Simon's model that cover the same range of exponents as the standard Simon model. The mathematical approach followed minimizes the amount of mathematical background needed for deriving the exponent, compared to previous approaches to the standard Simon's model. Reviewing what is known from other simple explanations of Zipf's law, we conclude there is no single radically simple explanation covering the whole range of variation of the exponent of Zipf's law in humans. The meaningfulness of Zipf's law for word frequencies remains an open question.

Keywords: Zipf's law, Simon model, intermittent silence.

INTRODUCTION

Zipf's law for word frequencies is one of the most striking statistical regularities found in human language. If f is the frequency of a word, the proportion of words having frequency f follows

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

where $\beta \approx 2$ in normal adult speakers (Zipf, 1932; Zipf, 1949; Ferrer i Cancho, 2005a). β is the exponent of Zipf's law. For simplicity, here we assume Eq. 1 for word frequencies, although other functional forms have been proposed (Chitashvili & Baayen, 1993; Tuldava, 1996; Naranan & Balasubrahmanyan, 1998). H. A. Simon proposed a process constructing a random text for explaining Zipf's law (Simon, 1955; Simon, 1957). At each iteration step, the text grows by one word. The (t+1)-th word will be either a new one (with probability ψ) or an old word (with probability 1-ψ). An old word means a word that has already appeared in the text. The old word is obtained choosing one word occurrence of the existent text sequence at random. All occurrences of words in the sequence have the same probability of being chosen. Equivalently, the old word can be chosen in the following way: the probability of choosing the word i is proportional to f_i, the normalized frequency of word i in the text sequence. The asymptotic distribution of the process follows Eq. 1 with

$$\beta = 1 + \frac{1}{1-\psi}, \tag{2}$$

[*] Address correspondence to Ramon Ferrer i Cancho, Dip. di Fisica, Università 'La Sapienza', Piazzale A. Moro 5, ROMA 00185, ITALY. E-mail: ramon@pil.phys.uniroma1.it.

(Simon, 1955; Simon, 1957; Rapoport, 1982; Manrubia & Zanette, 2002; Zanette & Montemurro, 2005).

Simon's model reproduces Zipf's law with $\beta > 2$. $\beta \approx 2$ is obtained for small values of ψ. Simon's model has been applied to many contexts. Some examples are the scale-free degree distribution of complex networks (Bornhodlt & Ebel, 2001) and the distribution of family names (Zanette & Manrubia, 2001; Manrubia & Zanette, 2002; Manrubia *et al*., 2003).

THE MODELS

Our models are simple generalizations of Simon's and follow essentially the same idea.

> **MODEL A.** At each iteration step, the text grows by m copies of a word ($m > 0$). The m copies will be either from a new one (with probability ψ) or from an old one (with probability $1-\psi$). The old word is obtained choosing one word of the text sequence at random, as in Simon's standard model. The probability of choosing the i-th word is proportional to f_i, the frequency of the word.

As it is formulated, model A means that the text sequence consists of blocks of m copies of the same word. That is not realistic when $m > 1$. A slight variation makes the generalized model more realistic, while giving the same word frequency distribution (see below for the mathematical details about the equivalence):

> **MODEL B**. At each iteration step, the text grows by one word. The $(t+1)$-th word will be either a new one (with probability ψ) or an old word (with probability $1-\psi$). The old word is obtained choosing one member of the text sequence at random. In this case, the probability of choosing the i-th word is proportional to k_i, the weight of the word. Every time a word is chosen, m is added to its weight. New words have zero weight. f_i , the frequency of the i-th word of the text is $f_i = k_i / m$.

When $m = 1$, we have $f_i = k_i$. In that case, Simon's model and model A and B are identical. The Appendix shows that the frequency distribution of models A and B follows Zipf's law (Eq. 1) and the exponent is given by Eq. 2 (as in the standard Simon model). Interestingly, the exponent does not depend on m ($m \geq 1$). The Appendix provides a calculation of $P(f)$ for Simon's model ($m = 1$) that is more detailed and requires less mathematical background than existing calculations (Simon, 1955; Simon, 1957; Rapoport, 1982; Manrubia & Zanette, 2002; Zanette & Montemurro, 2005).

DISCUSSION

Our extensions of Simon's model account for the same range of exponents as the standard Simon model. Neither the standard Simon model nor our generalizations cover the full interval of real exponents in word frequencies. As far as we know, real exponents lie within the interval [1.6,2.42] in single author text samples (Ferrer i Cancho, 2005a). $1 < \beta < 2$ has been found in some schizophrenics (Whitehorn & Zipf, 1943; Zipf, 1949; Piotrowski, 1995). $\beta = 1.6$ was reported by Piotrowski (1995). Young children have been shown to follow $\beta = 1.6$ (Brilluen, 1960; Piotrowski, 1995). $\beta = 1.7$ was found in military combat texts (Piotrowski,

1995). The standard Simon model and our extensions exclude the exponents of children and some schizophrenics.

Simon's model is a *birth stochastic process*. A 'birth' means choosing a word that has not yet been used. ψ is the birth rate. Simon's model has been extended to consider also 'deaths' (Manrubia & Zanette, 2002), which here means the possibility that a word occurrence disappears from the text. One may think that the fact that a word occurrence disappears means that the occurrence has been 'forgotten'. If ω is the death rate (i.e. the probability that a 'word' disappears at any iteration step), the extended Simon model by Manrubia and Zanette obeys Zipf's law (Eq. 1) with

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

When words never 'disappear', i.e. $\omega = 0$, Simon's standard model is recovered (Simon, 1955). It is easy to show that we have $\beta < 1$ or $\beta > 2$ for the birth-death model. When $1 - \omega - \psi < 0$, it follows that

$$\frac{1-\omega}{1-\omega-\psi} < 0, \tag{4}$$

which leads to $\beta<1$ using Eq. 3. $\beta<1$ is problematic since $P(f)$ can only be a probability function if time is finite. When $1-\omega-\psi>0$ it follows that $1-\omega>\psi$. So, assuming $\psi>0$ we obtain $1-\omega>1-\omega-\psi$. Dividing by $1-\omega-\psi$ on both sides of the previous equation we obtain

$$\frac{1-\omega}{1-\omega-\psi} > 1, \tag{5}$$

which leads to $\beta > 2$ using Eq. 3. The interval $\beta \in [1,2]$ is covered neither by the standard Simon's model, nor by our extension, nor by Manrubia and Zanette's extension.

While Manrubia & Zanette's extensions were not motivated by Zipf's law for word frequencies, Zanette & Montemurro extended Simon's model to make it more realistic for word frequencies (Zanette & Montemurro, 2005; Montemurro & Zanette, 2002). Recall N_t is the vocabulary size at time t. Their first extension (ZM1) takes into account that vocabulary growth in Simon's model is linear ($N_t \sim \psi t$) while real vocabulary growth is sublinear. ZM1 gets $N_t \sim t^{\nu}$ by replacing the constant rate ψ at which new words are added by a time dependant rate

$$\psi = \psi_0 t^{\nu-1}, \tag{6}$$

where $0 < \nu < 1$ and ψ_0 is the initial rate. ZM1 gives $\beta = 1 + \nu$ (Zanette & Montemurro, 2005). It is easy to see that $1 < \beta < 2$ for ZM1. The second extension (ZM2) tries to give recently introduced words more chance of being used again. After a series of simplifications, the final analytical model extends the standard Simon model with a probability γ. Any time an old word must be added, one word is chosen among all the words in the text regardless of its frequency (all existent words are equally likely) with probability and with probability $1-\gamma$, an old word is chosen as in the standard Simon model, i.e. with a probability proportional its frequency of occurrence. MF2 gives

$$\beta = 1 + \frac{1}{(1-\psi)(1-\gamma)}. \quad (7)$$

Simon's original model is recovered for $\gamma = 0$. Since $\gamma \geq 0$, it is easy to see that

$$\beta \geq 1 + \frac{1}{1-\psi}. \quad (8)$$

Knowing $\psi > 0$ we have $\beta > 2$. The third extension combines the extensions of ZM1 and ZM2 leading to

$$\beta = 1 + \frac{\nu}{1-\gamma}. \quad (9)$$

Knowing that $\gamma \geq 0$, we obtain $\beta > 1 + \nu$. Knowing $\nu > 0$ we get $\beta > 1$.

Simon's model and intermittent silence are among the simplest explanations for Zipf's law in word frequencies. Intermittent silence models consist of concatenating characters from a set including letter (or phonemes) and blanks (or silences). Every time a sequence a blank is produced, a new word starts (Miller, 1957; Miller & Chomsky, 1963; Mandelbrot, 1966; Li, 1992, Suzuki *et al.* 2005). For simplicity, it is assumed that all letters are equally likely. If σ is the probability of blank (or silence), intermittent silence reproduces Zipf's law with

$$\beta = \frac{1}{1 - \frac{\log(1-\sigma)}{\log L}} + 1. \quad (10)$$

The model in (Li, 1992; Suzuki *et al.*, 2005) is recovered when blanks have the same probability as any of the letters, that is, when $\sigma = 1/(L + 1)$. Assuming $L > 1$ and $\sigma > 0$, $L \to \infty$ and $\sigma \to 0$ give

$$-\infty < \frac{\log(1-\sigma)}{\log L} < 0. \quad (11)$$

Using the bounds in Eq. 11 on Eq. 10 we get $\beta \in (1,2)$. Therefore, intermittent silence accounts for a fraction of the interval of variation of real exponents. Interestingly, the widespread skepticism about the meaning of Zipf's law among scientists is mostly based on intermittent silence (Miller & Chomsky, 1963; Nowak *et al.*, 2000; Wolfram, 2002), which turns out to be incomplete.

The main argument against Zipf's law meaningfulness is that simple models can reproduce the law (Miller & Chomsky, 1963; Rapoport, 1982; Suzuki *et al*. 2005). It is worth to mention that Suzuki *et al.* take a special position, acknowledging the possible meaningfulness of Zipf's law for word frequencies but denying the relevance of Zipf's law for units (e.g. symbols) from unknown sources. Suzuki *et al.*'s main argument is that the presence of Zipf's law is not, in general, a sufficient for communication of any kind. If Zipf's law alone is considered a sufficient condition, false positives may be obtained. The criticisms mentioned above consider a very narrow interval of real exponents and use essentially two simple models, i.e. intermittent silence and the standard Simon's model. Those models are actually simple but do not cover, indeed, the whole range of real exponents.

We have seen that it is not easy to find a model covering the whole range of exponents. ZM3 is, as far as we know, the only modification of a simple model that covers the whole range. ZM3 is less simple than Simon's original simple model. In fact, arguing that ZM3 is a simple model is problematic, since it in incorporates two particular extensions at the same time. Therefore, one can safely, conclude that, so far, there is no radically simple explanation of Zipf's law covering the whole range of exponents.

The classic criticisms against Zipf's law meaningfulness need to be reviewed to the light of the range of real exponents. If the key problem against Zipf's law meaningfulness is finding a simple explanation, two different models are actually needed depending on the value of β: intermittent silence when $\beta < 2$ and Simon's model when $\beta > 2$. That is not very elegant since human language could be essentially the same system when considering the variations of β below or above 2 in normal adult speakers (Ferrer i Cancho, 2005a). Whether variations of β below or above 2 are the outcome of essentially the same communication system in normal adult speakers is a matter of current debate (Ferrer i Cancho, 2005a; Ferrer i Cancho, 2005b). We believe that ZM3 will be a crucial model in future discussions about the relevance of Zipf's law. In the absence of a radically simple model for Zipf's law covering the whole interval of real exponents, it seems wiser to give Zipf's law meaningfulness a chance in human language.

ACKNOWLEDGMENTS

This work was supported by 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.

APPENDIX

We start with the derivation of the word frequency distribution for model A. Thus, we consider an extended Simon where m identical words are added to the text at every time step t ($t > 0$). The identical words added are new with probability ψ or they are a copy of an existent word (which is chosen at random from the text) with probability 1 - ψ. The text sequences has N_0 different words and m_0 word occurrences at the 0-th step ($N_0 \leq m_0$). $P(f)$ can be easily derived with the mean field schema used for the Barabási-Albert scale-free network model (Barabási & Albert, 1999a, Barabási & Albert, 1999b). If k_i, the number of occurrences of the i-th word in Model A, and also t, are treated as continuous variables, then the expected variation of k_i is

$$\frac{dk_i}{dt} = (1-\psi)m\pi_i \ , \tag{12}$$

where

$$\pi_i = \frac{k_i}{\sum_{j=1}^{N_t} k_j} \tag{13}$$

and N_t is the number of different words at time t, π_i can be seen as a continuous rate of change of k_i.

Replacing Eq. 13 with $\sum_{j=1}^{N_t} k_j = m_0 + mt$ into Eq. 12 we obtain

$$\frac{dk_i}{dt} = \frac{(1-\psi)mk_i}{m_0 + mt}\,. \tag{14}$$

We define t_i as the time at which word i arrived. Integrating Eq. 14 for a word that appeared for the first time at $t = t_i$ with m copies we may write

$$\int_m^{k_i} \frac{dk_i}{k_i} = (1-\psi)m \int_{t_i}^{t} \frac{dt}{m_0 + mt}\,. \tag{15}$$

The previous Eq. leads to

$$k_i = m\left(\frac{m_0 + mt}{m_0 + mt_i}\right)^{1-\psi}\,. \tag{16}$$

According to Eq. 16., $\tau(k, t)$, the expected time of arrival of a word with k occurrences when the process is at time t (the time at which a word with k_i occurrences was added for the first time when the process is in the t-th iteration) becomes

$$\tau(k,t) = \frac{1}{m}\left((m_0 + mt)\left(\frac{k}{m}\right)^{-\frac{1}{1-\psi}} - m_0\right). \tag{17}$$

We define $P(k_i < k)$ as the probability that word i has less than k occurrences and $P(t_i > t)$ as the probability that word i arrives at time t or latter. We have that

$$P(k_i < k) = P(t_i > \tau(k,t))\,, \tag{18}$$

so

$$P(k_i < k) = 1 - P(t_i \le \tau(k,t))\,. \tag{19}$$

The number of words with $t_i < \tau(k,t)$ is $\psi\tau(k,t)$. Thus, the expected proportion of words with $t_i < \tau(k,t)$ is

$$P(t_i \le \tau(k,t)) = \frac{\psi\tau(k,t)}{m_0 + mt}\,, \tag{20}$$

where m_0+mt is the total amount of words at time t.

Replacing Eq. 19 with Eq. 20 into

$$P(k) = \frac{\partial P(k_i < k)}{\partial k} \tag{21}$$

we obtain

$$P(k) = \frac{\psi m^{\frac{1}{1-\psi}-1}}{1-\psi} k^{-1-\frac{1}{1-\psi}}. \tag{22}$$

Interestingly, the distribution does not depend on either t, m_0 or N_0 and the exponent depends only on ψ. For $m = 1$ as in the standard Simon model, Eq. 22 gives

$$P(k) = \frac{\psi}{1-\psi} k^{-1-\frac{1}{1-\psi}}. \tag{23}$$

Hence,

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

with

$$\beta = 1 + \frac{1}{1-\psi} \tag{25}$$

as in the standard Simon process. For deriving the word frequency distribution in model B, we need to take into account that $f = k/m$, where f is the random variable for the frequency of words and k is the random variable for the weight of words. We define $P(f)$ as the probability that a word has frequency f in model B. We have $\tilde{P}(f) = P(k(f))$ with $k(f) = mf$ so we get

$$\tilde{P}(f) \sim f^{-\beta} \tag{26}$$

using Eq. 24. Therefore, model B follows Zipf's law with the same exponent as model A.

REFERENCES

Barabási, A.-L. and Albert, R. (1999a). Emergence of scaling in random networks. *Science 286, 509-511.*

Barabási, A.-L., Albert, R., Hawoong, J. (1999b). Mean-field theory for scale-free random networks. *Physica A 272, 173-187.*

Bornholdt, S., Ebel, H. (2001). World wide web scaling exponent from Simon's 1955 model. *Physical Review E 64, 035104 (R).*

Brilluen, L. (1960). *Science and theory of information* (Russian translation). Moscow: Gosudarstvennoe Izdatel´stvo Fiz.-Mat. Literatury.

Chitashvili, R.J., Baayen, R. H. (1993). Word frequency distributions. In: G. Altmann and L. Hřebíček (eds.), *Quantitative text analysis: 54-135.* Trier: Wissenschaftlicher Verlag Trier.

Ferrer i Cancho, R. (2005a). The variation of Zipf's law in human language. *European Physical Journal B 44, 249-257.*

Ferrer i Cancho, R. (2005b). Zipf's law from a communicative phase transition. *Submitted to European Physical Journal B.*

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Transactions on Information Theory 38, 1842-1845.*

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In: P. F. Lazarsfield and N. W. Henry (eds.). *Readings in mathematical social sciences: 151-168.* Cambridge: MIT Press.

Manrubia, S.C, Derrida, B., Zanette, D.H. (2003). Genealogy in the era of genomics. *American Scientist 91, 158-165.*

Manrubia, S.C., Zanette, D.H. (2002). At the boundary between biological and cultural evolution: the origin of surname distributions. *Journal of Theoretical Biology 216, 461-477.*

Miller, G.A. (1957). Some effects of intermittent silence. *American Journal of Psychology 70, 311-314.*

Miller, G.A., Chomsky, N. (1963). Finitary models of language users. In Luce, R.D., Bush, R., & Galanter, E. (Eds.), *Handbook of Mathematical Psychology*. Vol 2. New York: Wiley.

Montemurro, M. A., Zanette, D. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics 4, 87-99.*

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

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

Piotrowski, R.G., Pashkovskii, V.E., Piotrowski, V.R. (1995). Psychiatric linguistics and automatic text processing. In: *Automatic Documentation and Mathematical Linguistics, 28(5):28-35.* First published in *Naučno-Techničeskaja Informacija, Serija 2, Vol. 28, No. 11. pp. 21-25, 1994.*

Rapoport, A. (1982). Zipf's law re-visited. *Quantitative Linguistics 16, 1-28.*

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Simon, H. A. (1957). *Models of Man.* Chapter 6: On a class of skew distributions functions. New York: John Wiley & Sons.

Suzuki, R., Tyack, P.L., Buch, J. (2005). The use of Zipf's law in animal communication analysis. *Animal Behavior 69, F9-F17.*

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

Whitehorn, J.C., Zipf, G.K. (1943). Schizophrenic language. *Archive of Neurology and Psychiatry 49, 831-851.*

Wolfram, S. (2002). *A new kind of science*. Champaign: Wolfram Media.

Zanette, D.H., Manrubia S.C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A 295, 1-8.*

Zanette, D.H., Montemurro, M.A. (2005). Dynamics of text generation with realistic Zipf distribution. *Journal of Quantitative Linguistics. In press.*

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge, MA: Harvard University Press.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Reading: Addison-Wesley.