Solvable null model for the distribution of word frequencies

J. F. Fontanari[1] and L. I. Perlovsky[2]
[1]*Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil*
[2]*Air Force Research Laboratory, 80 Scott Road, Hanscom Air Force Base, Massachusetts 01731, USA*
(Received 2 April 2004; published 25 October 2004)

Zipf's law asserts that in all natural languages the frequency of a word is inversely proportional to its rank. The significance, if any, of this result for language remains a mystery. Here we examine a null hypothesis for the distribution of word frequencies, a so-called discourse-triggered word choice model, which is based on the assumption that the more a word is used, the more likely it is to be used again. We argue that this model is equivalent to the neutral infinite-alleles model of population genetics and so the degeneracy of the different words composing a sample of text is given by the celebrated Ewens sampling formula [Theor. Pop. Biol. **3**, 87 (1972)], which we show to produce an exponential distribution of word frequencies.

DOI: 10.1103/PhysRevE.70.042901 PACS number(s): 87.10.+e, 02.50.Ey, 87.23.Kg, 89.70.+c

A remarkable aspect of natural languages is Zipf's law: if a large sample of words in a text are arranged in rank order, from most frequent to least frequent, then the dependence of the frequency P of a word on its rank ρ is very well described by the power-law distribution $P \propto 1/\rho$, regardless of the language or speaker [1]. The significance of Zipf's law in language, however, is obscure. On the one hand, the finding that texts produced by the random emission of symbols and spaces, so that words of the same length are equiprobable, also generate word frequency distributions that follow Zipf's law (more precisely, the generalized Zipf's law [2,3]) prompted the claim that this law is linguistically very shallow [4]. On the other hand, quantitative analyses of issues such as the evolution of syntactic communication [5] and the emergence of irregularities in language [6] usually assume that individuals use lexicons characterized by Zipfian word frequencies. Moreover, the fascinating enterprise of determining whether noncoding regions of DNA sequences have linguistic features, and hence whether they carry out biological information, is based on the assumption that Zipf's law is a crucial ingredient of language [7,8].

Clearly, if the random emission of symbols is accepted as the null hypothesis for the creation of texts in natural languages, then there is simply no need to seek explanation for Zipf's law, since it is accounted for by the null model. Not surprisingly, this viewpoint has been criticized on the grounds that a valid null model should be based on realistic assumptions on the factors that originate natural texts. In this vein, an alternative null hypothesis—the discourse-triggered word choice model—was put forward recently by Tullo and Hurford [9] (see also [10,11] for a similar proposal). In this setting, two sources of words are made available to speakers. The first is the environment, viewed as a large repository of distinct words, from where the speakers can choose words to start a conversation or to refresh an already worn out word store. The second is the words used in the preceding conversations, which leads to a positive feedback loop: the more frequent a word is, the more frequent it will become. According to those authors, this last and very plausible ingredient is responsible for the Zipfian distribution of word frequencies observed in natural texts. However, in this contribution we argue that this is not so. In particular, we point out that this word choice model is identical to the celebrated Wright-Fisher model of population genetics in the neutral regime [12] (see [13] for an inspiring overview) and then use the elegant mathematical apparatus developed in the early 1970s [14] to show that the word frequency distribution is exponential in the asymptotic limit.

The computer implementation of the discourse-triggered word choice model is as follows [9]. In the initial generation the word store is composed of N different words. Then N words are chosen randomly from this set, forming the word store of the second generation. Of course, some words of the original word store will be missing, while others will appear in several copies. The procedure is repeated with the new word store being selected from that of the previous generation, until the stationary regime is reached. The result of this procedure is, as expected, a drastic vocabulary loss—the final vocabulary being formed by a single word. (The term vocabulary refers to the set of different words in the word store.) To evade this problem, it is assumed that at each generation there is a probability that the selected word is chosen from the initial word store (i.e., the environment) rather than from the word store of the previous generation. This guarantees a continuous supply of new words, resulting in a nontrivial word distribution in the stationary regime. We emphasize that the outcome of this procedure is not the production of text or speech (to keep repeating a few words is not a good speech strategy), but the generation of a stationary word store characterized by a particular frequency spectrum, i.e., the average number of words that occur at a given frequency. Texts and discourses are then formed by drawing words at random from this word store.

The algorithmic procedure given above can be couched in a simple mathematical notation. Assume that the word-store size N is fixed and that there are K different words (the vocabulary size). Let m_i denote the number of copies of the ith word in the word store, so that its frequency is $x_i = m_i/N$. To avoid the uninteresting single-word vocabulary, let us assume that there is a mutation probability u per word so that mutation occurs to any of the other $K-1$ words with equal probability. Since we will eventually take the limits $N \to \infty$ and $K \to \infty$, this mutation scheme will always introduce a new word into the vocabulary. Strictly, the procedure for

vocabulary refreshment employed in the original model is equivalent to migration in population genetics, but if the population size and the number of allele types are large then it can be shown that migration and mutation play exactly the same role in the evolution dynamics [12]. Hence if a word i has frequency x_i at one generation, then the expected value of its frequency in the subsequent generation is $x_i'=x_i(1-u)+(1-x_i)v$ where $v=u/(K-1)$. The number of copies of word i at the next generation is then given by the binomial distribution

$$p(m_i)=\binom{N}{m_i}(x_i')^{m_i}(1-x_i')^{N-m_i}. \tag{1}$$

If one replaces the terms "word" by "allele" and "word store" by "population" then the Markov chain model just described is immediately identified with the Wright-Fisher model of population genetics in the neutral regime, i.e., in the case that there is no selection pressure favoring the choice of a particular word [12]. In the limits $N\gg 1$ and $u\ll 1$ such that the product $\theta=2Nu$ is finite, it can be shown that the probability density $\phi(x,t)$ that a given word, say l, has frequency x at time t obeys the Fokker-Planck equation [12]

$$\frac{\partial\phi}{\partial t}=\frac{1}{2N}\frac{\partial^2}{\partial x^2}[x(1-x)\phi]+\frac{\partial}{\partial x}\{[ux-v(1-x)]\phi\}. \tag{2}$$

The stationary distribution is then

$$\phi(x)=\frac{\Gamma[\theta K/(K-1)]}{\Gamma(\theta)\Gamma[\theta/(K-1)]}x^{\theta/(K-1)-1}(1-x)^{\theta-1}, \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. Our focus is on the limit $K\to\infty$, termed the infinite-alleles model in the molecular evolution literature [15], in which the number of distinct words (the vocabulary size) in the infinite word store is infinite too. Taking this limit in Eq. (3) yields $\phi\propto 1/K\to 0$, indicating that the particular word l we have considered will ultimately disappear from the vocabulary. This is expected since in this framework l can mutate to infinitely many different words but no word can mutate back to l. This feature is in conformity with estimates from glottochronology (i.e., the chronology of languages) that suggest the rule of thumb that languages replace about 20% of their basic vocabulary every 1000 years [16,17].

Although we cannot focus on the evolution of a particular word, we can calculate many other interesting properties of the word store. For instance, the mean number of words in the word store with frequency between x and $x+dx$ is simply

$$f(x)=\lim_{K\to\infty}[K\phi(x)]=\theta x^{-1}(1-x)^{\theta-1} \tag{4}$$

which can also be interpreted as the probability that a word occurs in the word store with frequency in $(x,x+dx)$ [14]. To illustrate the use of the "frequency spectrum" $f(x)$ let us write the frequencies of the various words occurring in the word store as $p_1,p_2,\ldots$ and consider any function of the form $\Sigma_i\psi(p_i)$ where $\psi(p_i)$ is of order of p_i^α with $\alpha\geqslant 1$. The expected value of any such function is then given by

$$\left\langle\sum_i\psi(p_i)\right\rangle=\int_0^1 dx\, f(x)\psi(x) \tag{5}$$

so that, in particular, $\langle\Sigma_i p_i\rangle=1$, as expected. For example, the probability that two words drawn at random are identical is $\langle\Sigma_i p_i^2\rangle=1/(1+\theta)$. Use of Eq. (5) allows the calculation of the probability π_k that in a random sample of n words drawn from the word store we find exactly k different words [14],

$$\pi_k=\frac{l_k\theta^k}{l_1\theta+l_2\theta^2+\cdots+l_n\theta^n} \tag{6}$$

where the coefficients l_i are the Stirling numbers of the first kind defined by $\theta(\theta+1)\cdots(\theta+n-1)=l_1\theta+l_2\theta^2+\cdots+l_n\theta^n$. Once it is known that the sample of n words contains k different words, we can address the question of how many times each word appears in the sample. The answer to this difficult question is provided by the celebrated Ewens sampling formula [14,18]

$$\mathrm{P}\{n_1,n_2,\ldots,n_k|k,n\}=\frac{n!}{k!\ l_k n_1 n_2\cdots n_k} \tag{7}$$

with $\Sigma_{i=1}^k n_i=n$. Here n_i is the number of copies of word i in the sample of n words.

Knowledge of Ewens formula allows us to obtain directly the distribution of word frequencies. To this end we need only to generate integers $n_1,\ldots,n_k$ with the probability distribution (7) and then sort them out according to their rank. More pointedly, if l is the most frequent word (i.e., n_l is the largest among the k integers) then n_l/n is the frequency of the first-rank word, and similarly for the words of lower ranks. This can be easily achieved using the Metropolis algorithm as follows (see, e.g., [19]). Consider n balls distributed among k urns, so that no urn is empty. The state of the system at step τ is specified by the vector $\mathbf{n}_\tau=(n_1,n_2,\ldots,n_k)$, where $n_i>0$ is interpreted now as the number of balls in urn i. Suppose the system is in state $\mathbf{n}_\tau$. We choose two distinct urns, say i and j, at random. Without loss of generality we consider urn i as the donor and urn j as the receptor. If $n_i=1$ then we maintain the current state in the next step, $\mathbf{n}_{\tau+1}=\mathbf{n}_\tau$. Otherwise we calculate the ratio

$$R=\frac{n_i n_j}{(n_i-1)(n_j+1)} \tag{8}$$

and move one ball from urn i to urn j if $R\geqslant 1$ so that the new state becomes $\mathbf{n}_{\tau+1}=(n_1,\ldots,n_i-1,\ldots,n_j+1,\ldots,n_k)$. On the other hand, if $R<1$ we generate a uniformly distributed random number r and move one ball from i to j provided that $R>r$, otherwise we keep the old state in the next step. These rules define the transition probabilities of a Markov chain, whose stationary state is distributed according to the Ewens sampling formula.

More specifically, we use the following procedure to produce the distributions of word frequencies. For fixed θ and n we generate the number of distinct words k according to distribution (6). Once k is known we can start the iteration of the Markov chain urn model just described. The initial state is chosen such that the n balls are distributed as uniformly as

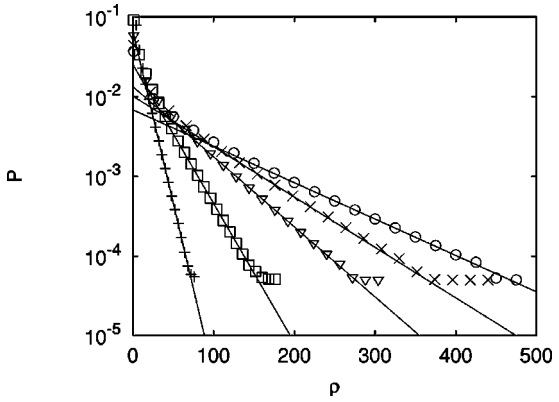

FIG. 1. Semilogarithmic plot of the average frequency P as function of rank ρ for $n=2\times10^4$ and $\theta=10$ (+), 25 (□), 50 (∇), 75 (×), and 100 (○). The straight lines are the numerical fitting obtained discarding the low rank region.

possible among the k urns. After a transient period of $1000n$ steps, we collect 1000 sample states at intervals of $100n$ steps. Sorting the k components of each state vector yields immediately the required information—the frequency as a function of the rank. The entire process is then repeated 50 times. At the end of this procedure we have 5×10^4 values for, say, the frequency of the word of rank 1. The results presented in Figs. 1–3 represent averages over these values.

The dependence of the average frequency of a word P on its rank ρ, depicted in a semilogarithmic scale in Fig. 1, reveals the exponential nature of the asymptotic regime of the word frequency distribution. This is corroborated by the results presented in Fig. 2, which show that the leveling off of the distribution in the high rank region is an effect of the finite sample size n. Moreover, that seems to be the sole effect of n, as indicated by the collapse of the data in the low and intermediate rank regions. A more quantitative perspective is provided by Fig. 3 where the rescaled frequency $P'=\theta P/0.62$ is plotted against the rescaled rank $\rho'=\rho/\theta$ in a semilogarithmic scale. The data for different values of θ (and n, as well) are fitted very well by the exponential function $P'=\exp(-\rho')$.

Our main result, namely, that the distribution of word frequencies of the discourse-triggered word choice model is exponential in the asymptotic regime, is in disagreement with the results of the proponents of the model, who found a power-law distribution [9]. The brute-force simulation method those authors resorted to, however, precludes a full statistical assay of the word frequency distribution. The study of the linguistic features of the noncoding DNA is another example where the finding of a power-law word frequency distribution is questionable [7,8].

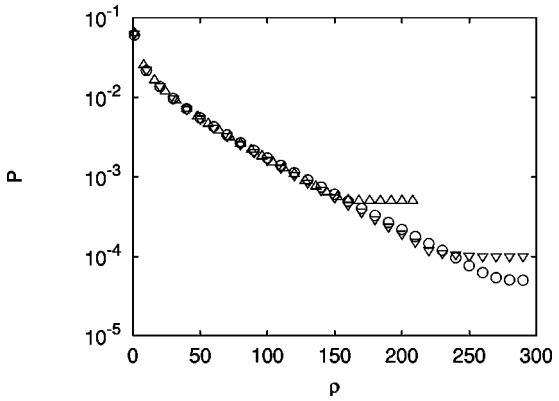

FIG. 2. Semilogarithmic plot of the average frequency P as function of rank ρ for $\theta=50$ and $n=2\times10^3$ (△), 10^4 (▽), and 2×10^4 (○).

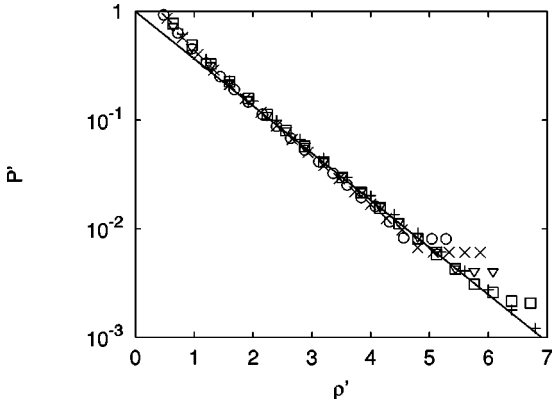

FIG. 3. Semilogarithmic plot of the rescaled average frequency P' as function of the rescaled rank ρ' for the data of Fig. 1. The straight line is the function $P'=\exp(-\rho')$.

From the standpoint of a null model, the failure of the discourse-triggered word choice model to reproduce Zipf's law is most welcome, as it calls for improvements of the basic model that may ultimately unveil the mechanisms responsible for the power-law distribution of word frequencies. In this line, we mention that a similar word choice model, in which the word store grows unconstrained from a single initial word and new words are generated by mutation, seems to exhibit a (nonstationary) power-law distribution of word frequencies [10,11]. Hence this model lacks the effective competition between words that results from the limitation of the word-store size. More importantly, the unbounded growth prevents the attainement of the key element of the present approach—a stationary, though not static, word store from where words are selected to form texts and discourses.

The notion that words compete and languages evolve similarly to individuals and populations was already familiar in Darwin's time [20]. The well-documented development of Romance languages from Latin (i.e., the gradual divergence of the languages of France, Italy, Spain, Portugal, and Romania from Latin, as well as from each other) offers a convincing proof that groups of related languages develop and diverge from a common ancestral tongue, similarly to gene lineages [16,17]. However, the use of analytical [21] and computational [22] methods derived from evolutionary biology to analyze language features and linguistic data is still incipient. The present contribution dovetails with these efforts by using the equivalence between the neutral evolution model and the discourse-triggered word choice model to calculate the distribution of word frequencies of this alternative null hypothesis for text production.

The research at São Carlos was supported by CNPq and FAPESP, Project No. 99/09644-9.

[1] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949).
[2] B. Mandelbrot, C. R. Hebd. Seances Acad. Sci. **232**, 1638 (1951).
[3] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).
[4] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982).
[5] M. A. Nowak, J. B. Plotkin, and V. A. A. Jansen, Nature (London) **404**, 495 (2000).
[6] S. Kirby. IEEE Trans. Evol. Comput. **5**, 102 (2001).
[7] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).
[8] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, J. Theor. Biol. **184**, 25 (1997).
[9] C. Tullo and J. R. Hurford (unpublished).
[10] L. Levitin, B. Schapiro, and L. I. Perlovsky, in *Proceedings of the Conference on Intelligent Systems and Semiotics '96* (National Institute of Standards and Technology, Gaithersburg, MD, 1996), Vol. 1, pp. 65–70.
[11] R. Günther, L. Levitin, B. Schapiro, and P. Wagner, Int. J. Theor. Phys. **35**, 395 (1996).
[12] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
[13] P. G. Higgs, Phys. Rev. E **51**, 95 (1995).
[14] W. J. Ewens, Theor Popul. Biol. **3**, 87 (1972).
[15] M. Kimura, Theor Popul. Biol. **2**, 174 (1971).
[16] C. Renfrew, *Archaeology and Language* (Cambridge University Press, Cambridge, England, 1987).
[17] J. P. Mallory, *In Search of the Indo-Europeans* (Thames & Hudson, London, 1989).
[18] S. Karlin and J. McGregor, Theor Popul. Biol. **3**, 113 (1972).
[19] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods* (Wiley & Sons, New York, 1964).
[20] G. Radick, Selection **3**, 7 (2002).
[21] M. A. Nowak, N. L. Komarova and P. Niyogi, Nature (London) **417**, 611 (2002).
[22] R. D. Gray and Q. D. Atkinson, Nature (London) **426**, 435 (2003).