ISSN 1066-369X, Russian Mathematics (Iz. VUZ), 2012, Vol. 56, No. 12, pp. 25–27. © Allerton Press, Inc., 2012.
Original Russian Text © V.V. Bochkarev and E.Yu. Lerner, 2012, published in Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika, 2012, No. 12, pp. 30–33.

The Zipf Law for Random Texts with Unequal Letter Probabilities and the Pascal Pyramid

V. V. Bochkarev* **and E. Yu. Lerner****

Kazan (Volga Region) Federal University, ul. Kremlyovskaya 18, Kazan, 420008 Russia

Received November 25, 2011

Abstract—We study a model of generating words with independent unequal probabilities of letters. We prove that the probability $p(r)$ of words of rank r has a power asymptotics. As distinct from a paper published earlier by B. Conrad and M. Mitzenmacher, we give a short proof, using elementary methods, and obtain an explicit formula for the degree of the power law.

DOI: 10.3103/S1066369X12120031

Keywords and phrases: *Zipf law, monkey model, order statistics, power laws, Pascal pyramid, recursive sequences, functional equations.*

As is known, the word "the" occurs in English texts more often than any other word; the second place (with respect to the frequency of occurrence) is occupied by the word "of"; thus, each wordform w in a text has its own *rank* $r(w)$, i.e., its number in the frequency table.

For a word w the *frequency* $f(w)$ is defined as the ratio of the number of occurrences of the word w in a text to the length of this text. According to the Zipf law (established in the first half of the last century), the product of $r(w)$ and $f(w)$ for an English text has an approximately constant value close to 0.1. At present time, owing to Google Labs, available are recognition results for 4% of all ever published books [1]. According to our calculations, the straight line constructed by the least square method from logarithmically transformed values of r and f (for convenience, we used the decimal logarithm) for most frequent (in 2000) English words takes the form $\log f = -1.05182 - 1.00026 \log r$. This agreement of the modern data (obtained for a small sample) with the Zipf law is the starting point of this paper.

The Zipf law can be explained in many ways; see [2] for a short review of relevant works by Russian researchers and the inference of this law from general properties of semiotic systems proposed by V. P. Maslov. In this paper we use a more traditional explanation of the Zipf law, namely, the "monkey" model. In this model, a monkey is assumed to type independently each time either a space (with probability p_0) or any of 26 English letters (with equal probabilities) [3–5]. A word is a sequence of letters between two spaces. Evidently, all words of the same length have equal probabilities, their ranks follow one by another. Denote by $p(r)$ the probability of obtaining a word of rank r. Evidently, $\exists c_1, c_2 : c_1 < \ln p(r) - \alpha \ln r < c_2$, where for $p_0 = 1/27$ we have $\alpha = \ln 27/\ln 26$. Here we do not take into consideration that in reality letters in a text have different frequencies; for the Russian language they approximately obey the law established by S. Gusein-Zade [6, 7] (probabilities are proportional to mean values of order statistics for the exponential distribution).

Rather recently, in the 12-page paper [8] B. Conrad and M. Mitzenmacher have generalized the monkey model for the case when letters have unequal probabilities. They have proved an inequality analogous to (1) (see below) by using the Tauber theorems for generating functions. In conclusion of the paper the authors write that it would be desirable to prove the power law for unequal probabilities without the use of these technical constructions and discuss possible approaches to such a proof. Here we propose a short proof; it was obtained in a different way, namely, with the help of properties of the Pascal pyramid. Note that in the case of the Markov dependence of probabilities of letters, the power law does not necessarily take place. In [9] we study the power and exponential laws for Markov chains and obtain explicit formulas for the corresponding parameters.

*E-mail: vbochkarev@mail.ru.
**E-mail: eduard.lerner@gmail.com.