Zipf's law, language and population genetics

José F. Fontanari
Instituto de Física de São Carlos
Universidade de São Paulo
Brazilfontanari@if.sc.usp.br

A remarkable feature of language is an empirical law made known by George K. Zipf in the early 1930s regarding the word frequency distribution in natural languages. In Zipf's analysis, one ranks the words in a text by decreasing frequency (i.e., number of occurrences). The resulting histogram is found to be linear on double logarithmic paper with a slope close to -1 for all languages studied. The origin of Zipf law, however, is still a unresolved issue. While some researchers argue that Zipf's law is connected to the hierarchical structure of languages, others claim that it is not a fundamental law since random texts exhibit the very same word frequency distribution. Recently, a proposal to explain the origin of this law in terms of simple principles of language use was put forward by researchers of the Language Evolution and Computation Research Unit of the University of Edinburgh. In particular, it is claimed that Zipf's law may be due to the fact that some words are used more frequently than others because language users hear them more frequently than other words (i.e., the more frequent a word is, the more frequent it will become). This idea was tested by those authors using a simple computer model. In the initial generation, the word-store of the language is represented by the set of integers ranging from 1 to K. Then a new word-store is generated by choosing randomly with replacement K number from this set. Clearly, some of the initial words will be absent and some will contribute with several copies to the new store. The procedure is repeated for a certain number of generations and the word frequency distributions calculated. To avoid the drastic reduction in the vocabulary size resulting from this procedure, an environmental factor is introduced in which there is a certain probability that a word comes from the environment (taken as the initial word-store) rather than from the word-store of the previous generation. As usual in this kind of analysis the numerical results are inconclusive, due perhaps to the small number of words K considered. In this contribution we argue that in the limit of infinite K, this model becomes identical to Kimura's infinite-alleles model of population genetics studied extensively in the 1960s. In particular, the reported drastic loss in vocabulary size is associated to the ultimate fate of an allele - fixation or extinction - in case the migration and mutation

pressures, which have a role similar to the stochastic environmental factor, are absent. If these pressures are taken into account then a stationary probability distribution for the words (or alleles) frequencies is shown to exist, being given by the celebrate Wright's formula (essentially, a beta distribution). Once this distribution is known, the frequency distribution of words in a finite sample taken from a text of infinite length can be obtained numerically with an arbitrary degree of accuracy. Ewens sampling theory (1972) provides the analytical framework to study this type of discrete multivariate distribution.