Available online at www.sciencedirect.com

SCIENCE DIRECT

Physica A 366 (2006) 495–502

PHYSICA A

www.elsevier.com/locate/physa

Statistical mechanical approach to human language

Kosmas Kosmidis*, Alkiviadis Kalampokis, Panos Argyrakis

Department of Physics, University of Thessaloniki, 54124 Thessaloniki, Greece

Received 27 September 2005
Available online 1 December 2005

Abstract

We use the formulation of equilibrium statistical mechanics in order to study some important characteristics of language. Using a simple expression for the Hamiltonian of a language system, which is directly implied by the Zipf law, we are able to explain several characteristic features of human language that seem completely unrelated, such as the universality of the Zipf exponent, the vocabulary size of children, the reduced communication abilities of people suffering from schizophrenia, etc. While several explanations are necessarily only qualitative at this stage, we have, nevertheless, been able to derive a formula for the vocabulary size of children as a function of age, which agrees rather well with experimental data.

Keywords: Language; Zipf law; Statistical physics; Language evolution

1. Introduction

Human language has recently become a subject of interdisciplinary character. Linguistic studies have traditionally been qualitative rather than quantitative. Recently, some attempts based on evolutionary game theory [1] have been made in an effort to understand language evolution, which have yielded some noticeable results. Particularly, interesting considerations were made in studies of competition between languages using mathematical [2] and computational models [3–10].

In this paper, we propose the assumption that human language can be described as a physical system within the framework of equilibrium statistical mechanics. Defining a Hamiltonian analogue that is associated with words, we are able to explain basic properties of spoken languages, such as the universality of the exponent of Zipf law [11], and to predict reasonably well the form of the curve for the vocabulary size versus age for young children. We, thus, demonstrate that statistical physics can provide an interesting formulation for the study of spoken languages and can unify aspects, such as the frequency distribution of words and the children's vocabulary learning rate, properties which at first glance seem completely different.

A rather remarkable feature, common to several languages is the so-called Zipf law [1], which states that if we assign the value $m = 1$ to the most frequent word of a language, $m = 2$ to the second one, etc., then the

*Corresponding author.
E-mail addresses: kosmas@kelifos.physics.auth.gr, kkosm@physics.auth.gr (K. Kosmidis), panos@physics.auth.gr (P. Argyrakis).

0378-4371/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.physa.2005.10.039

frequency of occurrence of a word with rank m is

$$C_m \sim m^{-a}. \tag{1}$$

This law has been verified experimentally for several languages with the exponent α value found to be universal and approximately equal to one. An alternative way, which is also used in the literature, to present Zipf law is to state that the proportion of words p_f whose frequency is f (taking values in the range 0–1) in a given sample text is modelled by a power function $p_f \sim f^{-\beta}$. The exponent β is related to the exponent α in Eq. (1) with the equation $1/\alpha = \beta - 1$. Although it is not immediately evident, the frequency–rank Zipf plot is equivalent to a plot of the cumulative distribution of p_f versus frequency f [12,13]. Ref. [12], in particular, contains a detailed proof of the above statement.

Traditionally, statistical mechanics does not deal with human language. It deals with physical systems, i.e., with collections of atoms, molecules or other elementary particles. According to statistical mechanics, when a system of particles is in equilibrium at constant temperature T, then it can be found in one of N states. The probability that it is found at a given state i with energy E_i is proportional to $\exp(-E_i/k_BT)$, the "Boltzmann factor." The temperature T is the "measure" of the interaction of the system with the environment.

2. The basic assumption of the model

Suppose that an individual possesses a vocabulary of N words. We treat the language department of the human brain as a physical system that can be found in one of N states. Each state represents one word. There is a one-to-one mapping between these states (which are enumerated using integers up to N) and words in the individual's vocabulary. If the system is found in state i, then the word associated with state i is pronounced. We denote as "temperature", T, a measure of the willingness (or ability) that the language speakers have in order to communicate. Common sense indicates that some words are more useful than others. The word "food" is essential and no organized group of people will go very far without it in their vocabulary. The word "heterogeneous" is probably not so useful since groups of people will probably survive without knowing it. Of course, usefulness is not only associated with meaning. For example, the word "and" is very useful because it is used to connect words. In this case, the usefulness originates from syntax rather than meaning.

Our *ansatz* is that the Hamiltonian of the system is $H(k) = \varepsilon \ln k$, where ε is a constant and k is a measure of a word's usefulness, i.e., we assign the value $k = 1$ to the most useful words, $k = 2$ to the second ones, etc.

Following the basic assumption of statistical mechanics, the probability to find a word with usefulness k is

$$p(k) = \frac{1}{Z} \exp\left(\frac{-H(k)}{k_BT}\right) = \frac{1}{Z} k^{-\varepsilon/k_BT}, \tag{2}$$

where Z is the partition function of the system [14]. Note, however, that Eq. (2) is not Zipf law. Zipf law connects the frequency of occurrence of words in a language with the rank of the word. Eq. (2) relates the probability of occurrence of a word with its usefulness k. Although we expect words of low k value to have a low rank m, and vice versa, there is no reason prohibiting two or more words from having the same value of k while these words have different rank.

3. Implications of the model

3.1. Divergence from the power-law at the initial part of a Zipf plot

It has been observed that while Eq. (1) is a straight line in log–log form, there are noticeable deviations in the early part of the line [15]. In Fig. 1, we plot the word occurrence, C_m, versus the word rank, m, using experimental data that come from a corpus (large collection of texts) consisting of publications in several Greek Internet sites up to May 2001, collected by Prof. Franz Guenthner at the University of Munich. It has been checked and used by T. Kyriacopoulou [16]. This corpus contains a total of about 2.6×10^7 words, about 2×10^5 different words and about 5.5×10^4 different lemmas. According to the Zipf law, such a frequency–rank plot should be a straight line in log–log plot as the word frequency distribution is a power-law). This is observed in Fig. 1, where the straight line is the best fit. Moreover, Eq. (2) implies that in the