Physics Letters A ••• (••••) •••–•••

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations

Ali Mehri [a], Maryam Jamaati [b]

[a] Department of Physics, Faculty of Science, Babol Noshirvani University of Technology, Babol, Iran
[b] Department of Physics, Iran University of Science and Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 7 March 2017
Received in revised form 28 May 2017
Accepted 29 May 2017
Available online xxxx
Communicated by C.R. Doering

Keywords:
Zipf's law
Long-rang correlation
Nonlinearity
Text mining
Human language

ABSTRACT

Zipf's law, as a power-law regularity, confirms long-range correlations between the elements in natural and artificial systems. In this article, this law is evaluated for one hundred live languages. We calculate Zipf's exponent for translations of the holy Bible to several languages, for this purpose. The results show that, the average of Zipf's exponent in studied texts is slightly above unity. All studied languages in some families have Zipf's exponent lower/higher than unity. It seems that geographical distribution impresses the communication between speakers of different languages in a language family, and affect similarity between their Zipf's exponent. The Bible has unique concept regardless of its language, but the discrepancy in grammatical rules and syntactic regularities in applying stop words to make sentences and imply a certain concept, lead to difference in Zipf's exponent for various languages.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Humans, as social beings, use language to exchange required information. Human language as a production of brain's cognitive ability, has a very complex structure to carry vast amount of information. Increasing the brain's volume and improving its structure in hominid evolution play a crucial role in language development [1]. Understanding the extraordinary structure and dynamics of human language, as a complex communication system, leads to uncovering the brain's function in thinking process. For this purpose, we need to find global features of language structure.

Despite a limited number of symbols, language has an unlimited capacity to express complex concepts. And its complexity is very high in comparison with the other communication systems. Therefore, we can take advantage of standard techniques for investigating the complex systems to study various features in language. Power-law regularities, as typical aspects of complex systems, have also been observed in human language [2]. Such a long tail distribution is hallmark of long-range correlations between the components of system. An important power-law discipline governing human language and many other complex systems, has been introduced by George Kingsley Zipf. He first proposed the principle of least effort for human behaviors. According to this principle, people act along the direction with probably minimal endeavor [3]. In this regard, when one can express an especial statement by a single word, applying several words to explain that concept will be unreasonable. From the physical viewpoint, the best performance will be achieved when the maximum value of information is transferred by expending minimum possible energy. Descending power-law relation between the number of occurrences (frequency) and the rank of frequency for language elements is considered as a clear effect of the least effort principle.

Zipf's law occurs in many natural and artificial systems, including many symbolic sequences (like ECG, EEG, MEG, genetic codes, etc.), which are applied in information coding and exchange. Adamic and Huberman observed the footprint of Zipf's law in different features of the cyberspace, e.g., level of routers transmitting data from one geographic location to another, the content of the World Wide Web, and the number of requests for webpages [4]. A simple extension of the Zipf analysis, so called *n*-Zipf analysis, has been applied to study correlations and biases in financial data [5,6]. It has attracted considerable interest from scientists in different research areas. For example, it is found that power-law models, like Zipf's law, can well describe the distribution of firm, city and many other man-made systems size distribution [7–10]. Such heavy-tail simple distributions for complex systems have been predicted by Simon's rich-gets-richer model [11]. He argued analytically that a population of flavored elements growing by either adding a novel element or randomly replicating an existing one would afford a distribution of group sizes with a power-law tail [12]. Barabási and Albert found a similar behavior in their growing network model [13]. Complex network theory, powered by such long tailed dis-

E-mail address: alimehri@nit.ac.ir (A. Mehri).

http://dx.doi.org/10.1016/j.physleta.2017.05.061
0375-9601/© 2017 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

tributions, has been applied to handle citation, collaboration and social networks [14–18]. It can also be used in language network analysis [19–21].

One of the first applications of Zipf's analysis in linguistics has been performed by Luhn for automatic keyword detection and abstract generation [22]. With a subtle glance at technical texts, one can find that typically each word type reflects only one meaning. It is very unlikely that authors apply several word types to express an especial concept. Moreover, the grammatical words, with low information content, are applied frequently in language. Therefore, Luhn sorted the word types of text according to their frequency, with regard to Zipf's method. Then he ignored the most frequent and the rarest words, and picked the middle ones as the relevant words. The distance between Zipf's plots can also be applied for authorship analysis. Havlin found that this distance between books written by the same author is smaller than the distance between books written by different authors [23]. In this regard, Bernhardsson et al. discovered that, for texts written by an author, a text with a certain length has the same Zipf's exponent as a text of the same length extracted from his/her imaginary complete infinite corpus [24]. It worth noting that skewness in the distribution of word intermittency and the average shortest paths have stronger correlation with writing style [25]. Empirical analysis on words' frequency indicates that, Zipf's exponent of the most popular keywords in top journals has completely different value in comparison with low impact factor journals [26].

Various evidences confirm the common origin of human languages [27]. More than 7000 languages are classified in over one hundred families [28]. A language family consists of a group of languages that have originated from a common ancestor. Zipf's law has been observed in many human languages, with different exponents depending on languages [29]. This work will focus on Zipf's law in one hundred live languages. We will extract Zipf's exponent for different translations of the holy Bible from 28 language families, and then we will compare them. We will also calculate average of Zipf's exponent for studied language families.

The organization of the remainder of the article is as follows. In section 2, we will briefly describe Zipf's law in natural languages. Section 3 contains a brief description about obtaining Zipf's exponent by fitting process. Then, in section 4 we will extract Zipf's exponent for the holy Bible translations in one hundred languages. Later, we will discuss the obtained results. Finally, in section 5, we will present a summary of the work.

2. Zipf's law

George Zipf noted the manifestation of several robust power-law distributions arising in different realms of human activity [3,30]. Among them, the most striking was the one referring to the word frequencies in human language [31,32].

One can sort the word types of language (a speech or a text) on the basis of their frequency. Thus, the most frequent word will place in the first position, and is assigned rank 1. The second most frequent word will appear in the second position, and is assigned rank 2 and so on. Finally, the rarest word type will be in the end of the list. Zipf claimed that, for all word types in the text, the multiplication of frequency and rank will be constant: $f(w)\,r(w) = C$. He grouped the words into three categories: a few numbers of frequent words, a median number of moderate words, and a large number of rare words. Pursuant to the principle of least effort, speaker/writer and listener/reader try to tolerate minimum endeavor. Hence, on the one hand, the author attempts to express his/her own opinion, by taking advantage of minimum number of frequent word types. On the other hand, the reader likes to remember large number of rare word types, for clear perception of author's objective. The competition between these two processes, leads to the such classification for language words, and power-law relation between words' occurrence frequency and rank:

$$f \propto r^{-\zeta}. \tag{1}$$

In this equation, r is called the frequency-rank of a word, and f is its frequency in a natural corpus. ζ is referred to as the Zipf's exponent, and its value is reported near unity for human written texts.

From an informative point of view, language words generally can be grouped within two major categories: content and function words. The function words serve to establish grammatical rules, and their frequency depends on the sentence structure. The head of Zipf's plot contains mainly function words, because a high percent of each text include such words. But the majority of words in the central part of the plot have a high information content. A subset of the content words are highly relevant to text subject and can serve as its keywords [33].

Many social, economic, biologic, etc. systems follow Zipf's law. All of them are composed of some elementary units, which are called tokens. These tokens can be grouped into larger entities, called types. Zipf's law deals with how tokens are distributed into types [34]. In the language case, types are the vocabularies which are distributed via a particular manner to express a special idea.

In quantitative linguistics phonemes/characters, morphemes, etc., may be considered as basic units of language [35]. But most studies have chosen the words as the elements, which are distributed with a particular pattern in order to express an special idea [36]. It is shown that various aspects of the complexity of a communication system may depend on the value of the Zipf's exponent [37]. Zipf's law is probably the most intriguing and at the same time well studied experimental law of quantitative linguistics, and it is extremely popular in its wider sense in the science of complex systems.

In addition to this well-known frequency-rank power-law discipline, Zipf has introduced some other empirical regularities in the realm of statistical linguistics [3]. In one of these laws, relationship between word types' frequency and number of their different meanings is explained. Authors tend to express all concepts by just a single word type, for sake of their ease. But, the readers prefer each word type implies a single distinct meaning. The interplay between these two opposite procedures with same strengths, leads to a power-law relation between the number of concepts (meanings) conveyed by a word type, m, and its occurrence frequency, f (or its frequency rank, r): $m \propto f^{1/2} \propto r^{-1/2}$. His another statistical law states that the keywords (content words) make clusters throughout the text. If d indicates distance between two successive occurrences of a word type, its frequency, n, will follow power-law function: $n \propto d^{-p}$, where the p exponent takes values near unity. He also found that, there is a reverse relation between length of a word type and its occurrence frequency in the text. Recently, it is shown that average information content is a much better predictor of word length than frequency. This indicates that human lexicons are efficiently structured for communication by taking into account interword statistical dependencies [38].

3. Zipf's exponent estimation

In practice, for Zipf's exponent extraction, words' frequency versus their frequency rank is sketched in a log-log plot. And then the linear part of the Zipf's plot, which is matched to power-law regime, should be fitted to a power-law model function to find the Zipf's exponent (ζ). In this work, we first calculate logarithm of relative frequency ($\ln(f/N_t)$), and logarithm of relative rank ($\ln(r/N_v)$) for all word types. N_t and N_v represent text length and its vocabulary size, respectively. Applying normalized frequency

UNCORRECTED PROOF

Please cite this article in press as: A. Mehri, M. Jamaati, Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations, Phys. Lett. A (2017), http://dx.doi.org/10.1016/j.physleta.2017.05.061

Download English Version:

https://daneshyari.com/en/article/5496198

Download Persian Version:

https://daneshyari.com/article/5496198

Daneshyari.com