ScienceDirect

Physics Letters A

Volume 381, Issue 31, 21 August 2017, Pages 2470-2477

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

Ali Mehri [a], Maryam Jamaati [b]

Show more

https://doi.org/10.1016/j.physleta.2017.05.061

Get rights and content

Highlights

- We check Zipf's law in hundred human languages.
- Zipf's exponent is extracted for hundred translation of Bible.
- All languages in some families have Zipf's exponent lower/higher than unity.
- Communications affect difference between Zipf's exponents of languages in a family.
- Distinct synthetic structures lead to different Zipf's exponents.

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.

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], [8], [9], [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 distributions, has been applied to handle citation, collaboration and social networks [14], [15], [16], [17], [18]. It can also be used in language network analysis [19], [20], [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.

Section snippets

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…

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.…

Zipf exponent for one hundred translations of the holy Bible

In many languages, words are separated by spaces or punctuation marks. In this work, the word is defined as any different string of characters between two whitespaces, after elimination of punctuation marks [51]. We extract Zipf's exponent for translations of the holy Bible to one hundred languages [52], by fitting their Zipf's plot with power-law model function.

Table 1 contains Zipf's exponent of the holy Bible translations to one hundred live languages, which is extracted by fitting. The fits …

Conclusion

Zipf's law is the most well-known empirical feature in quantitative linguistics. Since this law implies power-law behavior and long-range correlations, it is extremely helpful in its wider sense in the analysis of complex systems.

In this study, we compare this law in one hundred human languages. We extract Zipf's exponent for one hundred translations of the holy Bible. The Zipf's exponent for studied texts has values in $\zeta \in [0.765, 1.442]$ range. Average of Zipf's exponent for mentioned…

Recommended articles

References (59)

N. Vandewalle *et al.*
The *n*-Zipf analysis of financial data series and biased data series
Physica A (1999)

R. Hernández-Pérez *et al.*
Company size distribution for developing countries
Physica A (2006)

L. Gan *et al.*

Is the Zipf law spurious in explaining city-size distributions?
Econ. Lett. (2006)

S. Martinčić-Ipšić *et al.*
Multilayer network of language: a unified framework for structural analysis of linguistic subsystems
Physica A (2016)

A. Mehri *et al.*
The complex networks approach for authorship attribution of books
Physica A (2012)

S. Havlin
The distance between Zipf plots
Physica A (1995)

M.A. Montemurro
Beyond the Zipf–Mandelbrot law in quantitative linguistics
Physica A (2001)

W. Deng *et al.*
Rank-frequency relation for Chinese characters
Eur. Phys. J. B (2014)

T. Nabeshima *et al.*
Zipf's law in phonograms and Weibull distribution in ideograms: comparison of English with Japanese
Biosystems (2004)

J.M. Smith *et al.*
The Major Transitions in Evolution
(1997)

A. Mehri *et al.*
Power-law regularities in Human language
Eur. Phys. J. B (2016)

G. Zipf
Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology

(1949)

L.A. Adamic *et al.*
Zipf's law and the Internet
Glottometrics (2002)

H. Situngkir *et al.*
What Can We See from Investment Simulation Based on Generalized (m,2)-Zipf Law?
(2005)

B. Manaris *et al.*
Zipf's law, music classification, and aesthetics
Comput. Music J. (2006)

S.K. Baek *et al.*
Zipf's law unzipped
New J. Phys. (2011)

H.A. Simon
On class of skew distribution functions
Biometrika (1955)

P.S. Dodds *et al.*
Simon's fundamental rich-get-richer model entails a dominant first-mover advantage
Phys. Rev. E (2017)

A.L. Barabási *et al.*
Emergence of scaling in random networks
Science (1999)

B. Verspagen
Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research
Adv. Complex Syst. (2007)

D.R. Amancio *et al.*
On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks

Europhys. Lett. (2012)

G. Ahuja
Collaboration networks, structural holes, and innovation: a longitudinal study
Adm. Sci. Q. (2000)

S. Ghosh *et al.*
Understanding and combating link farming in the twitter social network

D.R. Amancio
Authorship recognition via fluctuation analysis of network topology and word intermittency
J. Stat. Mech. (2015)

S. Sisovic *et al.*
Comparison of the language networks from literature and blogs

H.P. Luhn
The automatic creation of literature abstracts
IBM J. Res. Dev. (1958)

S. Bernhardsson *et al.*
The meta book and size-dependent properties of written language
New J. Phys. (2009)

D.R. Amancio *et al.*
Comparing intermittency and network measurements of words and their dependence on authorship
New J. Phys. (2011)

Z.K. Zhang *et al.*
Empirical analysis on a keyword-based semantic system
Eur. Phys. J. B (2008)

There are more references available in the full text version of this article.

Cited by (42)

Quantifying relevance in learning and inference
2022, Physics Reports

Show abstract

Which words to teach: review and reflection

2022, International Encyclopedia of Education: Fourth Edition

Show abstract

Linguistic laws in biology

2022, Trends in Ecology and Evolution

> *Citation Excerpt :*
>
> …Rank and frequency are, by definition, negatively associated, and linguists studying this law (or the closely related modification, Zipf-Mandelbrot law, Table 1) focus on characterising the degree of linearity and steepness (exponent) of the log–log slope between these measures. In written English, the relationship is highly linear with an exponent ~1 [34] and a study of this law across 100 typologically diverse languages found a narrow exponent range centred around 1 but with some variation (0.76–1.44) [35]. Studies of Zipf's rank-frequency law beyond language come from different levels of biological organisation and diverse taxa (Figure 1C and Table 1)….

Show abstract

Power-laws in dog behavior may pave the way to predictive models: A pattern analysis study

2021, Heliyon

> *Citation Excerpt :*
>
> …Human languages have been known to demonstrate similar probabilistic distributions. Specifically, the frequency of occurrence of the words from a piece of text mostly follow a power-law distribution obeying the Zipf-Mandelbrot law [20]. Interestingly, this trend remains conserved and consistent across all known human languages and has been used to assign language-like property or lack thereof to other analogous datasets as diverse as inscriptions from ancient civilizations such as the Indus valley script [21], animal vocalizations like dolphin whistles [22] and bird songs [23], behavioral displays and sequences pertaining to courtship like the "push-up" displays of lizards [24] and even genetic distribution in an organism [25]….

Show abstract

Statistical metrics for languages classification: A case study of the Bible translations

2021, Chaos, Solitons and Fractals

Citation Excerpt :

…The competition between these two processes, leads to the such classification for language words, and power-law relation between words' occurrence frequency and rank. This statistical law was extensively studied in the context of quantitative linguistics [47,48]. An analogous behavior has been observed in a huge variety of other complex systems [49]….

Show abstract

On the emergence of Zipf 's law in music

2020, Physica A: Statistical Mechanics and its Applications

Show abstract

View all citing articles on Scopus ↗

View full text