




Chaos, Solitons & Fractals

Volume 144, March 2021, 110679

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

Ali Mehri  , [Maryam Jamaati](#)

[Show more](#) 

 Share  Cite

<https://doi.org/10.1016/j.chaos.2021.110679> 

[Get rights and content](#) 

Highlights

- Four statistical features are introduced for languages classification.
- We extract Zipf and Heaps' exponents, fractal dimension, entropy for 100 languages.
- Features have normal distribution around their mean value for Bible translations.
- Pearson correlation reveals the mutual connection between mentioned features.
- Cultural diversity can affect standard deviation of features over language families.

Abstract

Automatic language classification is an important contribution to linguistic research. Four statistical features concerning long-range correlations are applied to classify syntactic properties of languages. We calculate Zipf's exponent, Heaps' exponent, fractal dimension and entropy, for the Bible translations to one hundred live languages from twenty-eight language families. The Bible has unique concept regardless of its language, but the discrepancy in grammatical rules of the languages leads to difference in extracted measures from its various translations. The results show that, geographical distance and cultural differences can lead to statistical discrepancies. All extracted features for the Bible translations have normal distribution around their average value. This fact categorizes the languages into two groups; a majority of normal languages and a minority of abnormal ones. There is also evident (anti)correlation relation between each pair of the mentioned metrics due to their respective mechanism. Standard deviation of the considered statistical features over language families is affected by geographical distance between communities that speak to their languages and their cultural diversity.

Introduction

There are about seven thousand live spoken languages across the world, at least half of them are at risk of extinction [1]. They are grouped in several language families, according to their genealogical relations [2]. Various linguistic features can be exploited for languages classification. The distance between languages may depend on either their written or spoken form. The structural closeness of languages to each other has often been thought to be an important factor in foreign language learning. If the foreign language is structurally similar to the original language, it is claimed, learning should be easier than in cases where the foreign language is very different. This arises because of the complexity of languages, which differ by vocabulary, grammar, syntax, written form, etc. [3].

Many pioneer works have been done to find out various features of human language. Language, as a physical system, can be viewed from three different perspectives: microscopic, mesoscopic and macroscopic views. In the microscopic view, every word and its corresponding context contributes to the language grammar. On the other extreme, in the macroscopic perspective, a language can be characterized by a set of grammar rules and a vocabulary. From the mesoscopic viewpoint linguistic entities, such as the letters, words or phrases are the basic units and the grammar is an emergent property of the interactions among them [4]. As a modular system, language consists of a set of basic components, i.e. sentences, phrases, words, alphabets and etc. Different grammatical rules such as linguistic syntaxes discriminate the languages [5], [6]. Different syntactic, semantic, statistical, etc. properties can be applied by (un)supervised techniques, e.g., Naive Bayes classifiers,

artificial neural network, support-vector machines, k-nearest neighbors algorithms to reveal languages similarity [7], [8], [9]. Frequency distributions of n-grams and Levenshtein distances between words in the Swadesh list are widely used to quantify interlingual distances [10], [11], [12]. Pronunciation based methods will only be useful for checking similarity between languages that have close written or spoken forms. Hence, some other studies take advantage of statistical properties like information content (entropic measures) and analysis of linguistic networks for language clustering [13], [14].

It is generally known that, long-range correlations in symbolic sequences reveal their informational content [15]. Power-law regularities confirm the long-range correlation between elements in a system. Research has reported a large number of phenomena, such as the species within habitats, authors amongst scientific articles, actors within films, activation of genes, size of earthquakes, city populations, number of citations received by papers, sales of books, number of hits on webpages, etc. that follow the power-law regularities [16], [17], [18]. They also frequently emerge in many areas of linguistic organization [19], [20]. It is tempting to speculate that all distributions relating to language will approximate power-laws. Such distributions are characterized by a long tail consisting of a high proportion of very infrequent types [21].

Nature loves hierarchies, and the power-law is a statistical and emergent performance of hierarchies. In other words, the mechanism for the appearance of power law can be attributed to the existence of hierarchies in various natural and social phenomena. The power-law also possesses the characteristics of popularity and universality [22]. Scale invariance, suggested by the presence of power-laws, may be a natural consequence of criticality, either due to evolutionary tuning or self-organization without any outside influence [23]. Stochastic growth with a preferential attachment mechanism, eventually arrives at a power-law distribution of component abundances. This rich-gets-richer mechanism is at the basis of many stochastic models introduced to describe different component systems such as the Yule-Simon's model [24]. From the perspective of network theory, new connections(vertices) attach preferentially to nodes that are already well-connected [25]. It is proved that such networks evolve to a scale-free organization obeying a power-law distribution in which there is a long tail of nodes with low numbers of links and a small number of popular nodes with many links. Human language can be regarded as a growing network of interacting words with the small-world property as a result of natural optimization. At its birth, a new word interacts with several old ones. Interactions between words emerge from time to time, and new edges arise. Interestingly, the language network is also asymptotically scale-free due to its dynamic character with preferential linking [26].

The statistical analysis of linguistic laws shows long-range correlations between the constituents along the language [27], [28]. Zipf's law, as the most famous linguistic law, describes the power-law distribution of component frequencies [29]. This empirical law states that the frequency of a term has an almost inverse proportion to its rank in the frequency table. Another hallmark of long-range correlation in human language has been revealed by Harold Stanley Heaps. He introduced the sublinear scaling of the number of different component classes with system size [30]. The concept of fractality, which is introduced by Benoit Mandelbrot, can also be used to display long-range correlation in complex systems [31]. Fractal dimension is a ratio providing a statistical index of complexity comparing how in detail a pattern changes with the scale at which it is measured [32]. Words should be spread in a particular manner to convey a specified message. Such extraordinary distribution leads to long-range correlation between the words in all meaningful symbolic sequences [33], [34]. The amount of information conveyed by the language can be quantified by entropy [35], [36].

The quick and broad access to large databases of written text has crucial role in recent progressive applications of statistical methods in natural language processing. The studies provide insight on the mechanism of language production and linguistic relevance. We put emphasis on the quantification of the long-range correlations in the word ordering in several languages, and use them as appropriate criteria for comparison between languages. In this way, we apply four popular statistical quantities to study one hundred live languages from 28 language families. The initial conjecture is that, various grammatical structures of languages lead to extract different values of scaling exponents for them in Zipf's and Heaps' laws and their fractality. So, these three exponents along with entropy serve discrimination cues for languages classification.

The organization of the remainder of the article is as follows. In Section2, we will briefly introduce four applied statistical quantities including Zipf's and Heaps' exponents, entropy and fractal dimension. Then, in Section3 we will extract the mentioned quantities for one hundred translations of the Bible. Section4 provides average of the mentioned criteria for twenty-eight families, and detailed discussions about their obtained results. Finally, in Section5, we will present concluding remarks of the work.

Access through your organization

Check access to the full text by signing in through your organization.

Access through **your institution**

Section snippets

Statistical correlations in human language

The complex structure of human language enables us to exchange very complicated information. Its constituents interact with each other to form particular patterns according to rules and purposes. Such patterns represent the regarded meanings [37]. Some words, such as function words, appear in any topic, whereas content words only appear in the associated topics, suggesting different distributions. Extraordinary spatial distribution of words, that conveying a certain concept, results in...

Statistical criteria for one hundred translations of the Bible

We intend to extract the mentioned statistical criteria for the Bible translations to one hundred live languages. In this way, we use parallel corpus provided by Christodouloupoulos & Mark Steedman [57]. Electronic version of the applied data set is freely available online [58]. In pre-processing step, all characters with unit length are removed, and a string of symbols between two consecutive blank spaces is considered to be a word. Table3 shows the main statistics of all texts under...

Classification of the language families

In order to make a comparison between the language families, we calculate average of the four mentioned statistical metrics for the Bible translations to 28 studied language families. The average of Zipf's exponent, Heaps' exponent, entropy and fractal dimension for the families are listed in Table5. Their standard deviations are also reported there. Unfortunately, there exists just a single translation of the Bible for half of the families in our intended database. The standard deviation of...

Conclusion

In this contribution we have tried to use four statistical criteria for the purpose of language classification. Zipf's exponent, Heaps' exponent, fractal dimension and entropy, well-known empirical features in quantitative linguistics, have been calculated for one hundred translations of the Bible. All studied texts imply almost the same concepts, hence the difference in their extracted statistical measures comes from their different grammatical structures. The forenamed regularities belong to...

CRediT authorship contribution statement

Ali Mehri: Investigation, Conceptualization, Methodology, Software, Visualization, Writing - original draft. **Maryam Jamaati:** Conceptualization, Validation, Writing - review & editing....

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper....

Acknowledgement

The authors acknowledge the funding support of Babol Noshirvani University of Technology through Grant program No. BNUT/391023/99....

[Recommended articles](#)

References (61)

S. Wichmann

[On the power-law distribution of language family sizes](#)

J Linguist (2005)

G. Jäger

[Power laws and other heavy-tailed distributions in linguistic typology](#)

Adv Complex Syst (2012)

B. Bigi

[Using Kullback-Leibler distance for text categorization](#)

R. Rosenfeld

[A maximum entropy approach to adaptive statistical language modeling](#)

Comput Speech Lang (1996)

L. Lü *et al.*

[Deviation of Zipf's and heaps' laws in human languages with limited dictionary sizes](#)

Sci Rep (2013)

M. Ausloos

Generalized hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series

Phys Rev E (2012)

H.F. Jelinek *et al.*

Understanding fractal analysis? the case of fractal linguistics

Complexus (2006)

E. Najafi *et al.*

The fractal patterns of words in a text: a method for automatic keyword extraction

PLoS One (2015)

L. Rodgers *et al.*

Thirteen ways to look at the correlation coefficient

Am Statistician (1988)

P.K. Austin *et al.*

The Cambridge handbook of endangered languages (2011)

Ethnologue: languages of the world (22nd ed.)

B.R. Chiswick *et al.*

Linguistic distance: a quantitative measure of the distance between english and other languages

J Multiling MulticultDev (2005)

M. Choudhury *et al.*

The structure and dynamics of linguistic networks

C.H. Brown *et al.*

Automated classification of the world's languages: a description of the method and preliminary results

STUF-Lang Typology Univers (2008)

A. Mazzolini *et al.*

Zipf and heaps laws from dependency structures in component systems

Phys Rev E (2018)

J. Nerbonne *et al.*

Linguistic distances

Proceedings of the workshop on linguistic distances, LD'06, association for computational linguistics; Stroudsburg, PA, USA (2006)

A. Zubiaga *et al.*

TweetLID: a benchmark for tweet language identification

Lang Resour Eval (2016)

E. Asgari *et al.*

Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance

Proceedings of the workshop on multilingual and cross-lingual methods in NLP, San Diego, California (2016)

W.B. Cavnar *et al.*

N-gram-based text categorization

Proceedings of the third symposium on document analysis and information retrieval, Las Vegas, USA (1994)

D. Bakker *et al.*

Adding typology to lexicostatistics: a combined approach to language classification

Linguist Typology (2009)

F. Petroni *et al.*

Measures of lexical distance between languages

Physica A (2010)

J. Nerbonne *et al.*

Measuring dialect distance phonetically

Proceedings of the third meeting of the ACL special interest group in computational phonology (1997)

Y. Gao *et al.*

Comparison of directed and weighted co-occurrence networks of six languages

Physica A (2014)

E.G. Altmann *et al.*

On the origin of long-range correlations in texts

Proceedings of the national academy of sciences(2012)

M. Mitzenmacher

A brief history of generative models for power law and lognormal distributions

Internet Math (2004)

M.E.J. Newman

Power laws, Pareto distributions and Zipf's law

Contemp Phys (2005)

R. Sharman

Observational evidence for a statistical model of language

IBM UKSC report 205(1989)

T. Briscoe

Language learning, power laws, and sexual selection

Mind Soc (2008)

Yu S., Liang J., Liu H.. Existence of hierarchies and human's pursuit of top hierarchy lead to power-law. 2016....

T. Mora *et al.*

Are biological systems poised at criticality?

J Stat Phys (2011)

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

Cited by (3)

[A Unified Formulation for the Frequency Distribution of Word Frequencies using the Inverse Zipf's Law ↗](#)

2023, SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

[Extending Heaps' Law for Sublinear Vocabulary Growth on a Logarithmic Scale ↗](#)

2023, 31st IEEE Conference on Signal Processing and Communications Applications, SIU 2023

[Word synonym relationships for text analysis: A graph-based approach ↗](#)

2021, PLoS ONE

[View full text](#)

© 2021 Elsevier Ltd. All rights reserved.



All content on this site: Copyright © 2024 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the Creative Commons licensing terms apply.

