Zipf's law in Toki Pona

Dariusz Jan Skotarek
Institute of Applied Linguistics, University of Warsaw, Poland
https://doi.org/10.36505/ExLing-2020/11/0047/000462

Abstract

Zipf's Law states that within a given text the frequency of any word is inversely proportional to its rank in the frequency table of the words used in that text. It is a statistical regularity of a power law that occurs ubiquitously in language – so far every language that has been tested was found to display the Zipfian distribution. Toki Pona is an experimental artificial language spoken by hundreds of users. It is extremely minimalistic – its vocabulary consists of mere 120 words. A comparative statistical analysis of two parallel texts in French and Toki Pona showed that even a language of such scarce vocabulary adheres to Zipf's Law just like natural languages.

Keywords: Zipf's Law, Toki Pona, artificial languages, computational linguistics, statistics

Introduction

It is well-known that language users tend to choose certain words more often than others. George Kinsley Zipf discovered that it is a regular tendency, which can be accurately predicted with a mathematical formula. Zipf observed that having ranked the words in a text according to the number of their occurrences, different frequencies of ranked words create a harmonic series 1, 1/2, 1/3, 1/n. In other words, in a body of text of substantial volume the second most-frequent word will appear half of the times as the first most-frequent, third one will appear one third as often, etc. This means that the amount of times a word is used is proportional to 1/rank. This statistical method of language description turned out to be ubiquitous and accurate for all languages that have been tested so far – from English and Chinese to Esperanto and Meroitic. Zipfian regularity plotted on a logarithmic scale demonstrates a power-law distribution, as showed in Figure 1.

Toki Pona is an experimental philosophical artificial language created by Canadian linguist Sonja Lang in 2001. It is known amongst other constructed languages for its very limited vocabulary – its lexicon consists of only around 120 words. It is an extremely simple language. There are very few grammatical rules and no irregularities. In the light of its limited vocabulary, word formation takes place via combining elements, such as jan pona 'good person' meaning 'a friend'. It is a living language – there are hundreds of fluent Toki Pona speakers. It is gaining popularity – since the academic year 2020/2021 the University of Geneva will be offering a Toki Pona course.

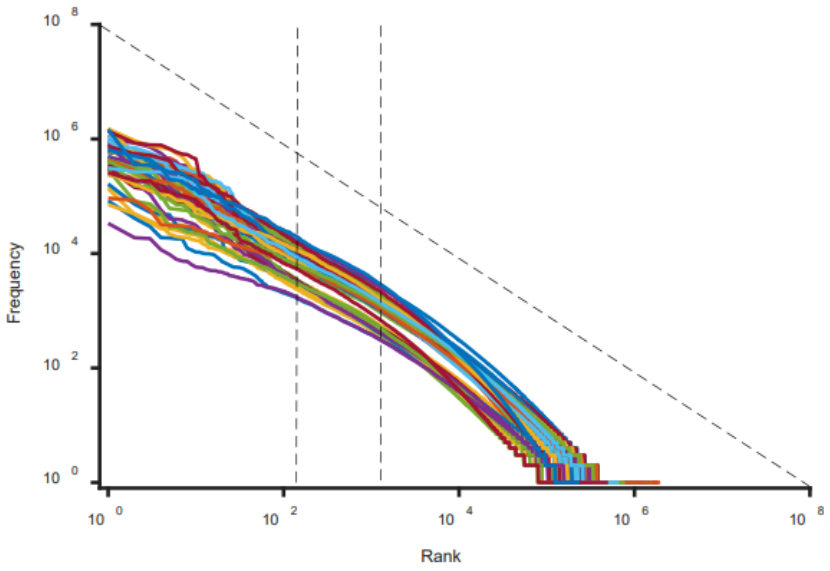

Figure 1. Frequency-rank distribution of 50 languages from various language families.

It is clear that Toki Pona differs tremendously from natural languages as well as from other artificial ones. Its vocabulary range of 120 words is nothing compared to that of, for example, English (approx. 250 000). If Zipf's Law is observed in languages with a wide vocabulary range, one could assume that Toki Pona could possibly "disobey" Zipf's Law. A natural language text consists of a handful of words that are used very often (i.e. articles and prepositions in English) but also of numerous hapax legomena – words used just once. This tendency makes natural languages "Zipfian". However, there are very few hapax legomena in Toki Pona – numerous words are used very often, since a given lexical unit constitutes a part of various lemmas. The point of the study was to "confront" Zipf's Law with Toki Pona.

Methodology

Two parallel texts were analysed – a legend "The life of Merlin" by Robert de Boron in the original, French version and its translation into Toki Pona done by a fluent Toki Pona speaker. Each text was parsed into a string of separate words. However, one aspect of the French language presented a problem in word-frequency analysis, namely elision, which is also marked in writing, for example l'ami 'friend' standing for le + ami. It raises a question whether one should view such cases as one or two words; an issue critical for a valid frequency-distribution analysis. In line with the Zipfian perception of what

constitutes a separate word, such instances were separated into two units. Parsed words were ranked according to the number of occurrences. Hence the rank-frequency distributions of the two texts were established.

Subsequently, those actual distributions were compared to those texts' perfect distributions predicted by Zipf's Law. The resulting correlation between the two values is a measure showing to what extent a given text obeys Zipf's Law.

Results

The French text's distribution represented a correlation with the Zipfian prediction of =0.89, with =1 being a perfect match. That result was in line with previous statistical analyses of natural languages. The Toki Pona text – contrary to possible initial assumptions – did obey Zipf's Law as well. Moreover, its distribution was coherent with the Zipfian prediction to the degree of =0.95, which is significantly higher than that of the French text.

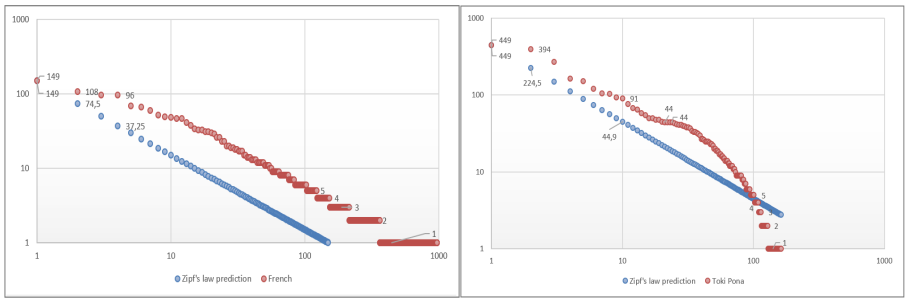

Figure 2. French (left) and Toki Pona (right) texts' lexical distributions and their respective Zipfian predictions.

Conclusions

The results of the study confirmed Zipf's Law's universality across both natural and artificial languages – even such unique, experimental ones like Toki Pona. It can constitute a valid argument in the discussion on possible explanations of Zipf's Law. One of such explanations is that every language is Zipfian because of its speakers who change it over time with their usage, which is guided by the principle of least effort. However, Toki Pona is a relatively young language, which did not undergo such user-driven change. Another explanation is that a Zipfian distribution is a reflection of a cognitive pattern that shapes human thinking – so that even while creating an artificial language its author unconsciously adhered to Zipf's Law. This thesis is supported by the fact that Sonja Lang confirmed that she was not aware that such law existed. Zipf's Law manifestations in phenomena other than language – such as city populations, web traffic or even size of Pluto craters – is a reason to assume that Zipf's Law is a purely statistical occurrence, with no deeper reason behind it. However, this

fact can also be used to pose a contrary argument – that Zipf's Law is a sign of something bigger, which eludes human understanding. Nevertheless, linguists shall test subsequent languages to see if indeed all human languages obey Zipf's Law, since – given the sheer quantity of languages on Earth – there are still many languages that remain untested in this respect.

References

Jiang, B., Yin, J., Liu, Q. 2015. Zipf's Law for All the Natural Cities around the World. International Journal of Geographical Information Science 29, 1–20.

Lang, S. 2014. Toki Pona: the language of good – the simple way of life. United States, Sonja Lang.

Reed, W.J., Hughes, B. D. 2002. From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. Physical Review E 66, 1–4.

Scholkmann, F. 2016. Power-Law Scaling of the Impact Crater Size-Frequency Distribution on Pluto. Progress in Physics 12(1), 26–29.

Smith, R. 2007. Investigation of Zipf-plot on the extinct Meroitic language. Glottometrics 15, 53–61.

Yu, S., Xu, C. Liu, H. 2018. Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. Semanticscholar.org.

Zipf, G.K. 1935. The Psycho-Biology of Language: An Introduction to Dynamic Philology. Cambridge, The MIT Press.

Zipf, G.K. 1949. Human Behavior and The Principle of Least Effort. Cambridge, Addison-Wesley Press.