Law of abbreviation

Alejandra Duque Maldonado & Guillem Bonet

April 2024

1 Introduction

One thing we can discover is that human language can be subject to statistical regularities.

One of the most studied statistical laws which affect linguistics is called the Zipf's law, a common phenomenon that seems to occur in complex systems where discrete units self-organize into groups, or types [1]. The law of abbreviation draws a relation between the frequency of appearance of a word and its length, as it generally states that the more frequent a word is, the shortest it tends to be. This is known as the "Principle of least effort", which was hypothesized as the result from pressure of having both: accurate and efficient communication in language.

As languages overall have a finite inventory that can be recombined into words, to satisfy these pressures languages assign shorter words to most frequent meanings and longer words to the less frequent.

For this task, we aim to determine if this law holds for languages of different families by measuring the lengths of the words, degree of optimality and correlation between the length and the frequency of words.

2 Materials and Methodology

In order to test whether the law of abbreviation holds in different languages, we picked 20 languages from 11 different families [1] for which we collected a parallel corpus. By using the NLTK library, we collected a parallel corpus with samples of the Universal Declaration of Human Rights (UDHR) for each of the languages we wanted to study. We used a parallel corpus so we could see the difference in results within different codes when treating with text of the same origin and context.

[1] We used the Glottolog database [2] for collecting these languages. More specifically: 6 Indo-european, 3 Austronesian, 2 Atlantic-congo, 2 Uralic, 1 Afro-asiatic, 1 Dravian, 1 Japonic, 1 Koreanic, 1 Sino-tibetan, 1 Turkic and 1 non-determined family (as for the case of Basque language).

We used the Python for gathering and pre-processing the texts and R for statistical analysis. Specifically, we used the spaCy library to tokenize the texts and obtain a table with token, length and frequency columns for each language. During this pre-processing, we also discarded all of the noisy data that resulted from this segmentation (i.e. punctuation marks, spaces and numerical data).

3 Results

Firstly, we looked at the correlation between length and frequency graphically by plotting 6 languages with length as a function of word frequency in Figure 1.

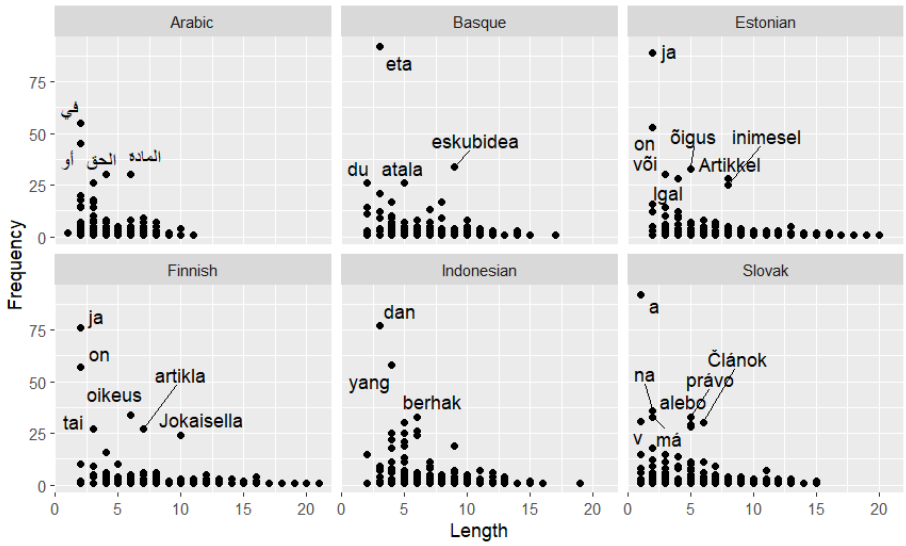

Figure 1: Length as a function of word frequency for Arabic, Basque, Estonian, Finish, Indonesian and Slovak.

Language	Family	Tokens	Types	Tau	p-value	C. p-value
Arabic	Austronesian	1318	725	-0.203	1.13e-10	1.02e-09
Basque	N/D	1236	652	-0.242	5.06e-14	8.09e-13
Bulgarian	Indo-European	2273	653	-0.229	3.46e-13	4.5e-12
Chinese	Sino-Tibetan	2693	532	-0.189	8.07e-07	4.84e-06
Estonian	Uralic	1250	654	-0.237	1.56e-13	2.19e-12
Finnish	Uralic	1113	672	-0.208	8.17e-11	8.17e-10
German	Indo-European	1330	545	-0.303	1.3e-18	2.59e-17
Indonesian	Austronesian	1302	488	-0.159	1.11e-05	3.33e-05
Irish	Indo-European	1640	598	-0.268	5.16e-16	9.79e-15
Japanese	Japonic	2325	517	-0.101	1e-02	2.01e-02
Kannada	Dravian	401	304	-0.0787	9.81e-02	9.81e-02
Korean	Koreanic	996	557	-0.176	2.84e-06	1.14e-05
Malay	Astronesian	1288	462	-0.177	2.16e-06	1.08e-05
Polish	Indo-European	1334	655	-0.238	7.46e-14	1.12e-12
Setswana	Atlantic-Congo	1735	459	-0.298	1e-15	1.8e-14
Slovak	Indo-European	1444	726	-0.235	1.53e-14	2.6e-13
Spanish	Indo-European	1559	509	-0.254	8.68e-13	1.04e-11
Tagalog	Afro-Asiatic	1509	453	-0.214	1.25e-08	9.96e-08
Turkish	Turkic	1310	696	-0.208	2.7e-11	2.97e-10
Yoruba	Atlantic-Congo	1437	385	-0.221	4.39e-08	3.07e-07

Table 1: Table displaying family, tokens, types, tau, p-value and corrected p-value for the 20 selected languages.

Also, in order to statistically check for the correlation, we ran the Kendall tau correlation test on the data for all the 20 selected languages and obtained the value of tau and p-value for each language. We also ran a Holm-Bonferroni correction on the p-values. This data is summarized in Table 1.

Language	Family	L_{min}	L	L_r	η	Ω
Arabic	Austronesian	4.06	4.67	5.42	0.87	0.553
Basque	N/D	5.77	6.75	7.99	0.855	0.558
Bulgarian	Indo-European	2.83	3.57	6.17	0.791	0.777
Chinese	Sino-Tibetan	1.01	1.01	1.05	1	1
Estonian	Uralic	5.83	6.7	8.43	0.869	0.664
Finnish	Uralic	6.83	7.68	9.3	0.889	0.656
German	Indo-European	5.17	6.29	8.5	0.822	0.665
Indonesian	Austronesian	5.08	6.47	7.75	0.785	0.48
Irish	Indo-European	3.81	4.87	6.85	0.782	0.651
Japanese	Japonic	1.29	1.62	1.94	0.797	0.493
Kannada	Dravian	6.93	7.8	8.33	0.888	0.375
Korean	Koreanic	2.4	2.89	3.24	0.829	0.413
Malay	Astronesian	5.13	6.49	7.69	0.791	0.47
Polish	Indo-European	5.18	6.26	8.04	0.828	0.624
Setswana	Atlantic-Congo	3.46	4.59	7.31	0.755	0.707
Slovak	Indo-European	4.67	5.65	7.23	0.827	0.617
Spanish	Indo-European	4.02	5.18	7.6	0.776	0.675
Tagalog	Afro-Asiatic	4.23	5.42	7.98	0.78	0.682
Turkish	Turkic	5.5	6.4	7.63	0.86	0.578
Yoruba	Atlantic-Congo	2.73	4.03	4.93	0.677	0.408

Table 2: Table displaying family, tokens, types, mean word length (L), the minimum baseline (L_{min}), the random baseline (L_r), the degree of optimality (η) and the optimality score (Ω) for the 20 selected languages.

After the statistical tests, we also computed other metrics which can be useful to identify patterns and better understand our results. We calculated the mean word length (L), the minimum baseline (L_{min}) and the random baseline (L_r). After obtaining the previous values, we also computed the degree of optimality (η) and the optimality score (Ω). We summarized this data in Table 2.

With the data represented in Table 2, we created a plot which shows L_r as function of L in order to visualize it in Figure 2. We also included L_{min} encoded as the size of the dots to add a third dimension to the data and visualize L, L_r and L_{min} at the same time.

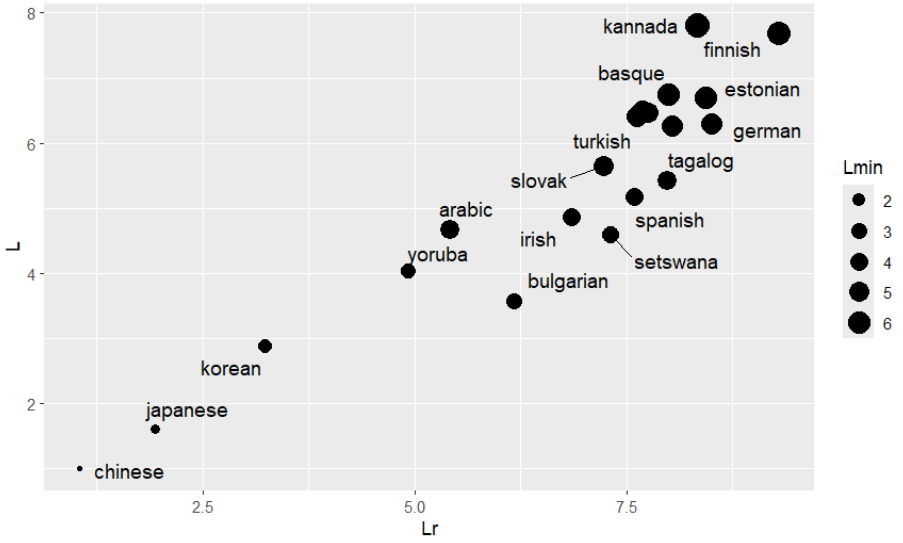

Figure 2: L_r as function of L with L_{min} encoded as dots size for the 20 selected languages.

4 Discussion

In terms of the length of the words, we see that for all of the languages we tested the average length tends to be between 1 and 8 characters. The minimum length of words tends to be one to two characters less than the mean length of words for each of the languages, meaning that the minimum and the average length of the words in all observed cases tend to be proportional.

In Figure 2, we see how the languages that are more compressed are the ones which have in common the fact that they come from different parts of Asia. The most compressed of them seems to be the case of Chinese, which could be explained by the fact that Chinese is a monosyllabic language; meaning that the words in this language have complex structures which are able to represent whole concepts [3]. However, for the case of Japanese and Korean, which are polysyllabic languages (i.e. words can have multiple syllables), according to our results they might reflect a tendency to have words that consist of monosyllables in its system. In contrast, within the least compressed languages, we see languages like Kannada, which seem to have a tendency for words to be larger when compared to the rest of observed languages. What this tells

is that in monosyllabic languages as Chinese, the system allows the information of words to be compressed in a single complex character, in comparison to least compressed languages, where information is encoded in longer strings of characters.

When looking at optimization data in the results, we see that the degree of optimality values obtained ranged between the scores of 0.6 and 1. On the other hand, when measuring the optimality scores, we saw these values ranged from 0.3 to 1.

In [4] it is mentioned that real languages are optimized in an average of 30%. However, we find that for our parallel corpus results reflected much higher values which ranged between 60% to 100% in coding efficiency. We could assume the reason why these values are higher is due to the size of our corpus, as for this task we used a relatively short text that contained a variety of types. Furthermore, this text is of a specific genre and does not represent the language as a whole.

Regarding the correlation, we found very small p-values (smaller than 0.05 in all cases) and, therefore, we did find correlation between frequency and length of tokens. All the tau values are negative with different values between -0.3 and -0.1, which means that there is a negative association between frequency and length; therefore the bigger the frequency, the smaller the length. This results are aligned with the law of abbreviation. While all results showed a strong correlation, the tau values did differ, which means that some languages have a more negative correlation than others.

5 Conclusion

In conclusion, from the data we were able to verify that the analyzed languages follow the law of abbreviation. While we found differences in the the correlations, all of them were negative. Also, we found that some languages are more compressed than others, and that the compression seems to improve the optimization of languages.

The main takeaway from this analysis is that, regardless of differences among languages, the law of abbreviation appears to be universal as described in [5].

References

[1] Corral and I. Serra, "The brevity law as a scaling law, and a possible origin of zipf's law for word frequencies," *Entropy*, vol. 22, no. 2, p. 224, Feb. 2020. [Online]. Available: http://dx.doi.org/10.3390/e22020224

[2] Glottolog, "Glottolog 5.0 - — glottolog.org," https://glottolog.org/, [Accessed 11-04-2024].

[3] A. Michaud, "Monosyllabicization: patterns of evolution in asian languages," *Monosyllables: from phonology to typology*, pp. 115–130, 2012.

[4] R. Ferrer-I-Cancho and C. Bentz, "The evolution of optimized language in the light of standard information theory," in *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2018. [Online]. Available: http://dx.doi.org/10.12775/3991-1.029

[5] C. Bentz and R. Ferrer-i Cancho, "Zipf's law of abbreviation as a language universal," 2016.