Lab 2. Zipf's frequency laws and the theory of power laws

Alejandra Duque Maldonado, Giacomini Nicolò

May 6, 2024

1 Introduction

In this laboratory we are going to use the Zipf's frequency law and Zipf's number-frequency law in order to obtain valuable analysis about the structure and dynamics of natural language, and to go into detail in various applications across linguistic research and computational linguistics. For this analysis we made use of data extracted from a parallel corpus of different languages and language families in order to calculate the exponents of the different power laws and contrast if these calculations were approximately consistent with given definitions of these laws.

2 Methods

In this analysis, to see the relation between power laws we focused on retrieving the exponents for the values of α, β and β' by doing logarithmic scales and performing robust linear regressions. α represents the exponent of the power law in the context of Zipf's rank-frequency law, which describes the relationship between the frequency of a word and its rank in a given language. β represents the exponent of the power law in the context of the rank spectrum, which is used to analyze the relationship between frequency and rank. β' is the exponent of the power law in the backward cumulative number of types that have a frequency equal to or smaller than a given frequency. Once these values were retrieved, we contrasted how the values obtained could be related to the equations shown in (1). Once these exponents were retrieved, we were able to plot graphs showing how the frequency-rank law shown in six of our samples, in addition to a smoothed out version of this.

2.1 Data description

For this analysis, we made use of data used for the previous exercise, which consists of extracted data from a parallel corpus of the Universal Declaration of Human Rights (UDHR). Within these corpus we collect data from 20 languages from 11 different families according to the Glottolog database [1]. In summary, we collected data from: 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 (i.e the case of Basque language).

From these previously collected data, we used python libraries to retrieve and pre-process the texts. Firstly, we retrieved our paralell corpus from the Natural Language Toolkit (NLTK) library corpus, which contained all the translations of the Universal Declaration of Human Rights. Secondly, we tokenized these texts using the respective tokenizer models for each language, which are part of the sPacy libraries.

For this exercise, we extracted from each text the number of tokens (i.e. all of the words) and the types (i.e. unique occurences of the words). With this data we were able to obtain the ranks in relation to the frequency of the occurrences and, additionally, the spectrum of lexical diversity within the text (i.e. how many different words are contained in the text).

2.2 Results

The results are these:

Language	**Family**	**Tokens**	**Types**	α	β	β'
Arabic	Austronesian	1318	725	-0.68	-0.51	-1.53
Basque	N/D	1236	652	-0.68	-0.55	-1.54
Bulgarian	Indo-European	2273	653	-0.93	-0.76	-1.18
Chinese	Sino-Tibetan	2693	532	-0.73	-1.05	-0.90
Estonian	Uralic	1250	654	-0.77	-0.50	-1.58
Finnish	Uralic	1113	672	-0.75	-0.36	-1.82
German	Indo-European	1330	545	-0.73	-0.70	-1.28
Indonesian	Austronesian	1302	488	-0.73	-0.76	-1.24
Irish	Indo-European	1640	598	-0.81	-0.74	-1.19
Japanese	Japonic	2325	517	-0.93	-0.98	-0.97
Kannada	Dravian	401	304	-0.43	-0.34	-2.16
Korean	Koreanic	996	557	-0.66	-0.48	-1.56
Malay	Austronesian	1288	462	-0.70	-0.80	-1.17
Polish	Indo-European	1334	655	-0.73	-0.57	-1.45
Setswana	Atlantic-Congo	1735	459	-0.91	-0.90	-1.04
Slovak	Indo-European	1444	726	-0.76	-0.55	-1.47
Spanish	Indo-European	1559	509	-0.87	-0.78	-1.16
Tagalog	Afro-Asiatic	1509	453	-0.93	-0.82	-1.10
Turkish	Turkic	1310	696	-0.64	-0.56	-1.55
Yoruba	Atlantic-Congo	1437	385	-0.68	-1.01	-0.92

Table 1: Values of α, β and β'

The relationships that should be satisfied theoretically are the following:

$$\begin{aligned} \beta &= \frac{1}{\alpha} + 1 \\ \beta' &= \frac{1}{\alpha} \\ \beta &= \beta' + 1 \end{aligned} \tag{1}$$

By the table we can see that in most of cases the relationship are not respected. The only cases where the relationships are approximately valid are the following: Arabic, Basque, Korean, Polish, Setswana, Slovak, and Turkish.

If the theoretical relationship are not respected, this could mean that the language is complex and exhibits various patterns and irregularities. It could indicate the presence of additional factors influencing word frequency distributions that Zipf's laws does not take into account. Another reason is due to the inaccuracies or biases in the data used for analysis that can lead to deviations from the expected relationships.

In the following pictures we can see the results of six languages:

To generate these graph, we used a robust linear regression, in particular Theil-Sen's method that allows to ignore outline data. The relation that we analyzed is $log(freq)$ $log(ranks)$. The graphs outlines that more the rank is high more the frequency of that words is low. In the plots we can see that many words are has rank equal to zero, this because the scale that we used is a logarithmic scale. Therefore, the relation that we can see the graphs follows an exponential curve.

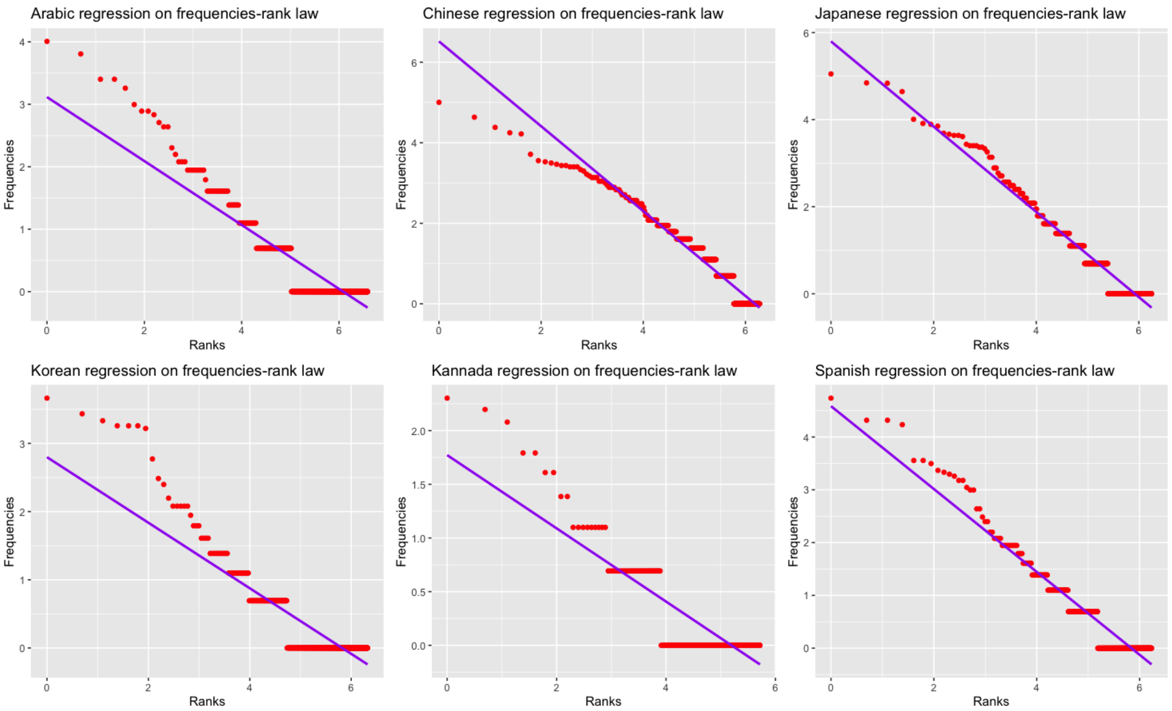

Figure 1: Regression with Theil-Sen's method.

In the last figure, we can see the smoothed curve using a multiplicative binning:

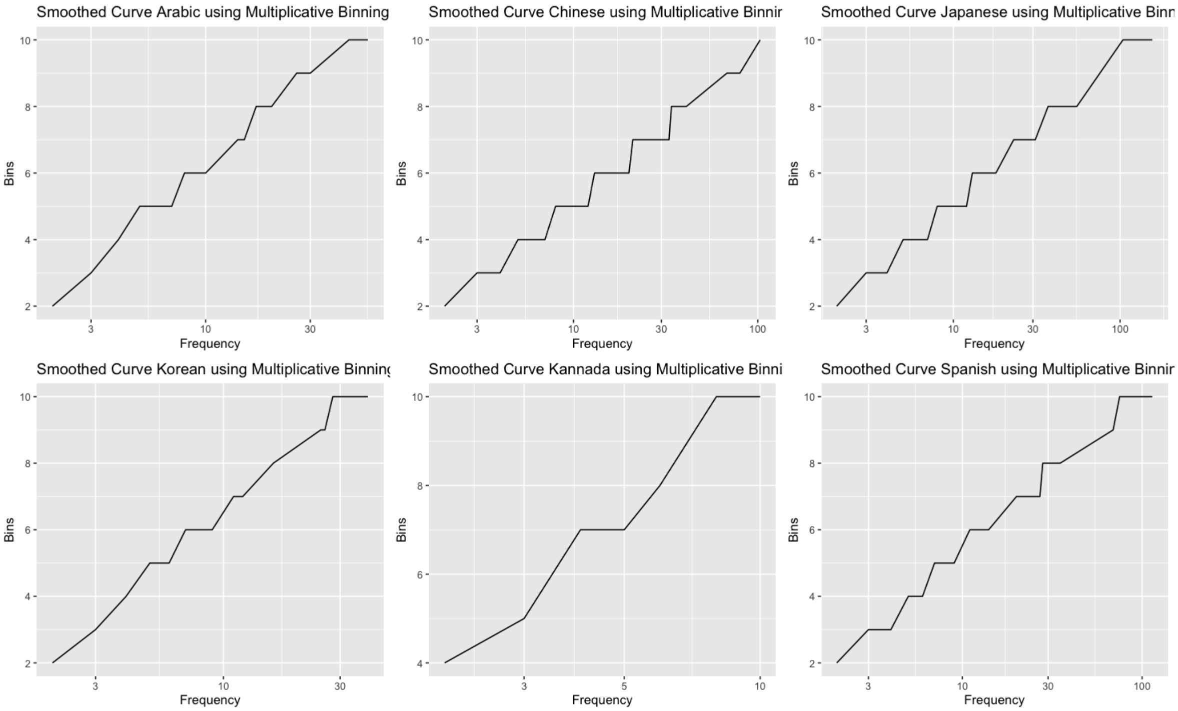

Figure 2: Smoothed Curve using Multiplicative Binning.

In the graphs, the bins represent intervals or ranges into which the frequencies of words are grouped. These intervals are created based on the frequency values of the words in the dataset. It is possible to notice that more the frequency is higher, more the number of bins is high. It is interesting to notice that also in this case, the relation is exponential. This is showed by the fact that the scaled used are logarithmic, therefore as the size of the text increases, the proportion of unique words decreases, this lead to a greater portion of words used more than once, with frequency grater than 1. Additionally, less common words become less frequent while more common words become more frequent.

3 Discussion

Within our results, even when in all languages we did not found a consistent coherence with the equations and the approximations, we can still see from some of the plots that there are some regularities that can be observed. These regularities are observed in the distribution of the words in both, in terms of the frequency and the rank.

As we know, words that have higher frequency in a text are placed in higher ranks (e.g. rank 1) and less frequent words to be placed in lower ranks (e.g. rank 6). In terms of how the types of words are distributed in terms of the frequency, we can observe how we tend to have a higher density towards the lower ranks of our samples, and a lower density around the higher ranks. From these data, it would be possible for us to be able to perform an analysis on the lexical diversity within a text, and how it can relate to the frequency of appearance of words within a text.

From our observations in Figure 1 and Figure 2, we could see how the words with the higher frequency tend to have less diversity than the words that have lower frequencies, meaning that many of these words might as well be unique occurrences. What this tells us is, that lexical diversity seems to manifest in the form of unique occurrences to less frequent words. In other words, the most frequently used words are also the most redundant ones, whereas the more diverse lexicon tends to be less redundant and has a tendency to only be seen as an unique occurance.

4 Conclusion

The spectral distribution of word frequencies is a very important aspect of understanding the lexical structure of any text and natural language. In addition to the rank distribution, it provides a quantitative-systematic perspective on the composition of words within a text. This frequency changes dynamically with the size of the text, and it follows an irregular law, wherein an increase in text size leads to decrease the frequency of utilization of rare words, including those used only once.

The connection between the lexical frequency and the rank distribution is evident in the analytical descriptions using various functions. Many analytical descriptions of the frequency spectrum are based on Zipf's law, with the conclusion that both rank and frequency distributions can be described by the a power function. This function shows the principle of proportionality of the relative decrease or increase based on the variables. This is known as the law of constant relative growth.

Mathematical models are approximations and may not perfectly fit empirical data. However, these models provide valuable insights into the underlying structure of linguistic nature.

In quantitative linguistics and science in general, it's important to consider the general form of data presentation and their value is based on the comparison between theoretical and empirical distributions. Even though our results may not align perfectly with theory, they describe clearly the phenomena that we wanted to study and thanks to this experience we can showing a realistic outcome.

References

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

[2] J. Tuldava. The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, 3(1):38–50, April 1996.