

Verification of Zipf's law of abbreviation in two languages: English and Spanish Under the Context of News Report

WANG SHUNKAI, YANGJIAHUI
CCiL

I. Introduction

George Kingsley Zipf (1972) made a classic observation about the relationship between a word's length and its frequency; the more frequent a word is, the shorter it tends to be. He (1972) hypothesized that this universal design feature arises as a result of individuals optimizing form-meaning mappings under competing pressures to communicate accurately but also efficiently. In the present work, we test two languages: English and Spanish under the context of the news report about the earthquake. English text contains 822 and the Spanish text, 516. We choose these two languages based on their widespread popularity in the world and these two languages are both formed by Latin scripts. Therefore it's easier for us to conduct the statistical result. When it comes to the genre selected, it is intuitive that news are usually delivered with formal use of language, for which less non-regular forms are expected. A negative correlation between word frequency and word length is found in both types of languages. The result is in line with Zipf's original proposal.

II. Material and Methods

The two texts selected as the sample are news report of the earthquake happened in Turkey. The English one is retrieved from BBC (2023) and the Spanish one, from *EL País* (2023). Considering their similarity in genre and topic, it is presumed that the variables are, to some degree, controlled so that we may focus on traits of the languages *per se*.

The word frequencies are counted as the number of occurrences of lemmatized and lowered **word tokens**. With *a priori* intuition that Spanish text usually has little clitics and that there are more compound words in English, we take different strategies of word filtration. For Spanish text, we resort to *spacy* with the model *es_core_news_lg* to automatically accomplish the tokenization and then only keep the tokens composed of letters. For English text, *RegexTokenizer* is applied with a manually made rule to identify as many compound words as possible. Given some single letters stranded after tokenization, we take another step to remove them except *a* and *i*, which are considered typical English words.

Our first presumption is that the law is always followed so that all the words bearing a frequency higher than any shorter ones will be categorized as abnormal. In this way, 48% of the words in English text and 34% in Spanish text are labeled deviant (see the part of abnormal words in our code). To better present the , we appeal then to scatters and fitted curve to visualize the data. Basically, what we construct is a 2D graph with length on x-axis and frequency on y-axis and the distribution of each word token will be presented by spots. Finally, a fitted curve will show the tendency. Additionally, to verify the generality of our texts, another analysis is taken with two corpus: *cess_esp* (Spanish) and *brown* (English). We extract **word types** instead of tokens and count how many words types are of a certain length. The result is presented by bar charts.

III Result and Discussion

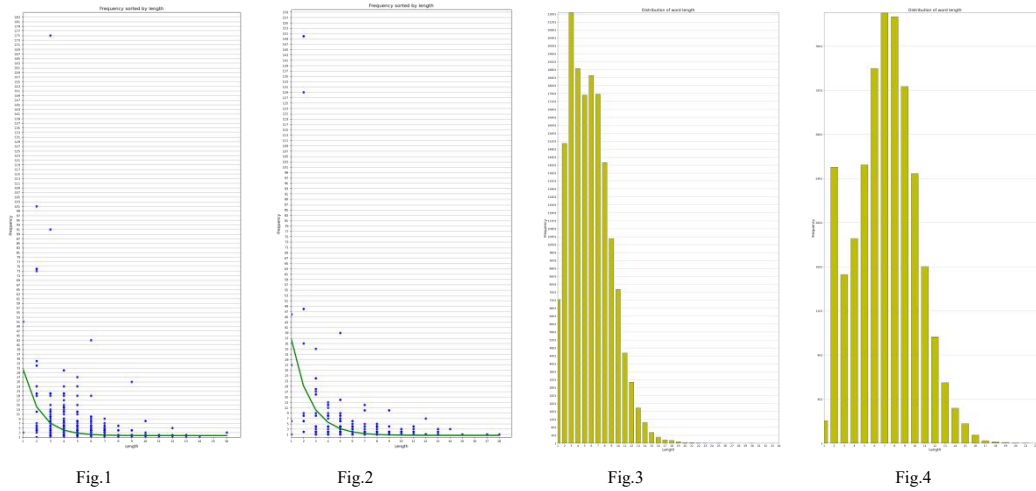


Fig. 1. The relationship between frequency of words tokens (y-axis) and number of word characters (x-axis) for English text

Fig. 2. The relationship between frequency of words tokens (y-axis) and number of word characters (x-axis) for Spanish text

Fig. 3. The relationship between frequency of words types (y-axis) and number of word characters (x-axis) for English corpus

Fig. 4. The relationship between frequency of words types (y-axis) and number of word characters (x-axis) for Spanish corpus

The magnitude of words stands in an inverse relationship to the occurrence of tokens. We believe that this result is very unlikely to occur by chance. In general, the law is embodied by the curves in both cases and we observe no overt discrepancy by comparing the result from two languages, which is conforms with our presumption. However, in regard with specific words, not every single word necessarily appears less frequently than shorter ones. In other words, Zipf's brevity law should be taken from a macro perspective. Otherwise, we would come to the deviant conclusion that most words violate it (as our first assumption). Also, there are some words with abnormally high occurrence in both texts with length of 2-3 characters. By printing them, we find that they belong to stopwords. A common strategy for text cleaning in NLP tasks consists in removing the stopwords before further processing. In this sense, Zipf's law can be applied for identifying potential stopwords (Fenrilli and Izzi, 2021). Last but not least, the distribution of words (visualized by density of spots per x-value in figure 1 and 2) is mostly consistent with the word type distribution in corpus), suggesting that they are representative. Hence, we may cogently predict that the same tendency applies to other news data.

Although the data is visualized, we carried out comparison procedure mainly based on observation of graphs, which degrades the result's objectivism. To simplified word filtration, some compound words in Spanish may be deliberately removed. For their richness in vocabulary, we take the corpora as lexical source, but they are no match for dictionaries. As for the data, it would be more persuasive to involve different types of texts since lexical preference varies from genres (e.g. singular first-person pronouns avoided in thesis).




Finally, despite the result, we may argue that it is because of the proximity of the two languages that we achieve highly similar outcomes. A large-scale research conducted on 1262 texts and 986 different languages also revealed the negative correlation but with some cases not as typical as others (Bentz and Ferrer Cancho, 2016). Their graphs shows more mild slope in language system like Sino-Tibetan. In ideographic languages the syllables and length don't account for the complexity and it is also difficult to segment words. Overall, there is no denying that Zipf's Law of Abbreviation is universal, but the extent to which it fits depends on distinct languages.

Reference:

- Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics* (pp. 1-4). University of Tübingen.
- Ferilli, S., Izzi, G. L., & Franza, T. (2021). Automatic Stopwords Identification from Very Small Corpora. In *Intelligent Systems in Industrial Applications* (pp. 31-46). Springer International Publishing.
- Kingsley, Z. G. (1972). *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner Publishing Company.
- Mourenza, A. (2023, February 7). *Turquía y Siria buscan contra reloj supervivientes de los terremotos que han causado más de 7.800 muertos* <https://elpais.com/internacional/2023-02-07/turquia-y-siria-buscan-contra-reloj-supervivientes-del-terremoto-que-ha-causado-5000-muertos.html>
- BBC. (2023, February 8). <https://www.bbc.com/news/live/64533954>

Code:

Our code can be accessed through <https://github.com/YANG-JIAHUI99/NLP-Zipf-s-law.git>

	Ex1-English.ipynb
	Ex1-Spanish.ipynb
	Ex1_mixed.ipynb

Claim of contribution:

This group work is conducted by 2 persons, WANG SHUNKAI and YANG JIAHUI. Shunkai Wang's responsibility is to convert the two texts into the code as well as extract number information about the co-occurrences of words with different lengths by means of python programming, and at last generate the images which help us give the conclusion. Jiahui Yang's responsibility is to write the paper work.