María Leva Jiménez and Eloi Singla i Milian
Natural Language Processing

Zipf's Law of Abbreviation two corpora comparison

Introduction

Zipf's Law of Abbreviation establishes that there is an inversely proportional relationship between a word's length and its frequency; that is, the shorter a word is, the more frequent it usually is. This law has been claimed to be a universal property of language (Kanwal et al., 2017). This report will examine two languages, English and Scottish Gaelic, to see whether both of them conform to this law, at least in a specific textual genre, news reports. The expectation is that both of them will, and that results for both languages will be similar, due to the rule being universal, at least in theory. Scottish Gaelic has been selected due to personal interest. Its specific particularities (such as the presence of very long words) might affect its compliance to Zipf's law to an unknown degree. On the other hand, English has been selected on the basis of availability of resources. Finally, the genre was selected based on the ones that were present in the Gaelic corpus (explained below) and were not narrations.

Material and methods

In order to answer this report's question, we have used two corpora: the Brown corpus for English and the ARCOSG corpus for Scottish Gaelic. The Brown corpus (Francis and Kučera, 1961) is an electronic corpus available in the NLTK library. It features different textual genres, among which journalistic texts like news can be found. The ARCOSG (Annotated Reference Corpus of Scottish Gaelic) (Lamb et al., 2020), available at GitHub, also contains different types of texts, both oral and written. The texts selected for this analysis were news scripts.

To preprocess the data, the text samples from both corpora were cleaned, and appended to lists. Both corpora were already tokenized, so we only had to separate them from the tags. For the ARCOSG corpus, we did it by separating the word-tag pairs first, and then eliminating the words tagged as punctuation while keeping only the rest of the words. Since, for reasons unknown, some punctuation still remained, we got rid of it by comparing every token against the list of punctuation symbols from the library "string". Finally, all tokens were lowercase and added to a list.

For the Brown corpus, we got the raw text via the NLTK library, using the "category" parameter to select only news. Tags were eliminated from the raw string using regular expressions. The text was tokenized using the NLTK library, and added to a list after lowercasing them and deleting punctuation marks (using the aforementioned list) and numeric characters. The reason why tags were utilized in the first case and not in the second is that ARCOSG had a "general punctuation" tag (with further specifications), whereas Brown has a different one for each symbol.

To answer the main question, the function calculate_zipf() was defined to calculate word frequencies and aggregate them based on word length, fit a linear regression, make a plot showing the data, and return different calculated values. With these, the two corpora were compared by means of another function, compare_corpora(), which included the Kolmogorov-Smirnov 2-sample test to compare the word length-frequency distributions of both corpora.

Results

Overall, both language samples of these specific textual genres conform to the Law (Figures 1 and 2, see Appendix), a result which meets our initial expectations. According to our results, Zipf's law affects English to a higher degree (cf. R^2= 0.93 and slope=-0.20 for English and R^2= 0.91 and slope=-0.16 for Gaelic)[1], even though the difference between the languages is minimal (Kolmogorov–Smirnov test=0.0741)[2].

Looking at the plots, however, one can see that the shorter words, those of length 1, were not conforming to the Law. We then tested the samples excluding these. In English they were first-person singular pronouns and indefinite determiners, while in Gaelic they were some prepositions and particles. The law still holds for both languages (Figures 3 and 4, see Appendix), and indeed they seem to follow it more closely than in the test will all the words (cf. R^2= 0.97 and slope=-0.21 for English and R^2= 0.94 and slope=-0.12 for Gaelic). The difference between both corpora is also smaller (Kolmogorov–Smirnov test=0.0516). A table comparing the results of all four samples can be seen in the Appendix.

These differences might have to do with the limitations of our data, which lie precisely in the particular stylistic features of this textual genre. The more direct, brief, telegraphic style of the chosen texts may condition the kind of words that appear in the samples (e.g. for English, first-person-singular pronouns are not usually typical on news unless there are quotations). It would be interesting to analyse other textual genres to see how the results might change (e.g. literary texts that make use of flowery language and less frequent words). Perhaps comparing different journalistic texts (e.g. news reports vs editorials or opinion columns) would also yield different results, assuming in these texts there is a higher occurrence of first-person singular personal pronouns and more creative uses of language. In any case, we believe any differences between textual genres in these languages would be minimal, since these results point to the law still holding for them.

[1] R^2 measures the strength of the linear fit. A higher value means that there is a stronger correlation between word length and frequency. The value of the slope of the linear regression measures the effect of Zipf's law: the more negative the value, the stronger the effect.

[2] The Kolmogorov–Smirnov test value represents how much both corpora differ in their distribution. A higher value means there is a bigger difference between the corpora.

Appendix

Figure 1: Linear regression plot for English

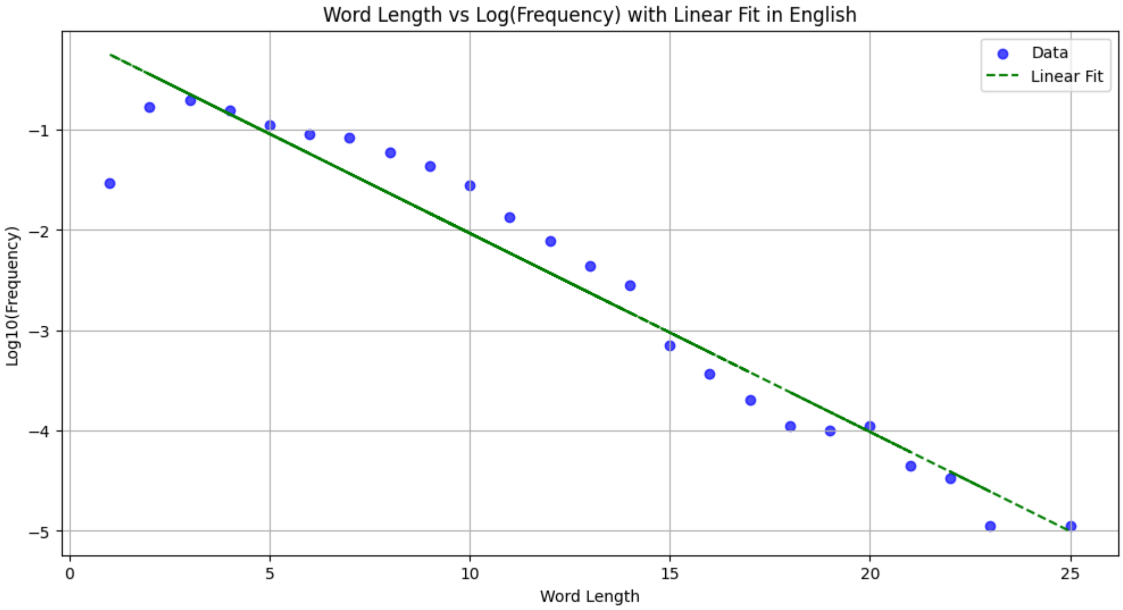

Figure 2: Linear regression plot for Scottish Gaelic

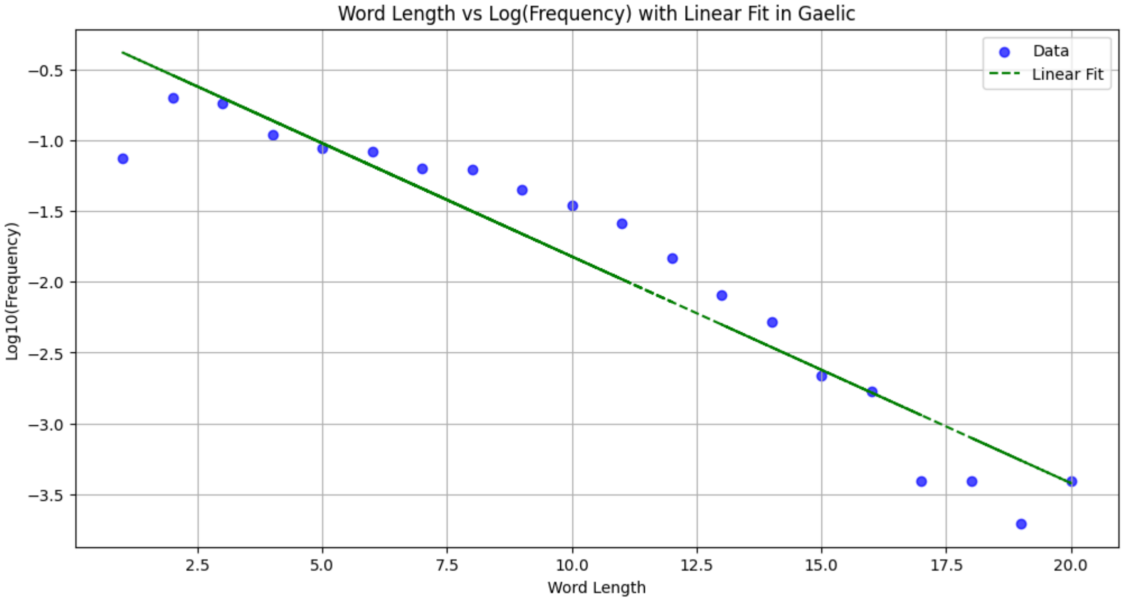

Figure 3: Linear regression plot for English, excluding words of length = 1

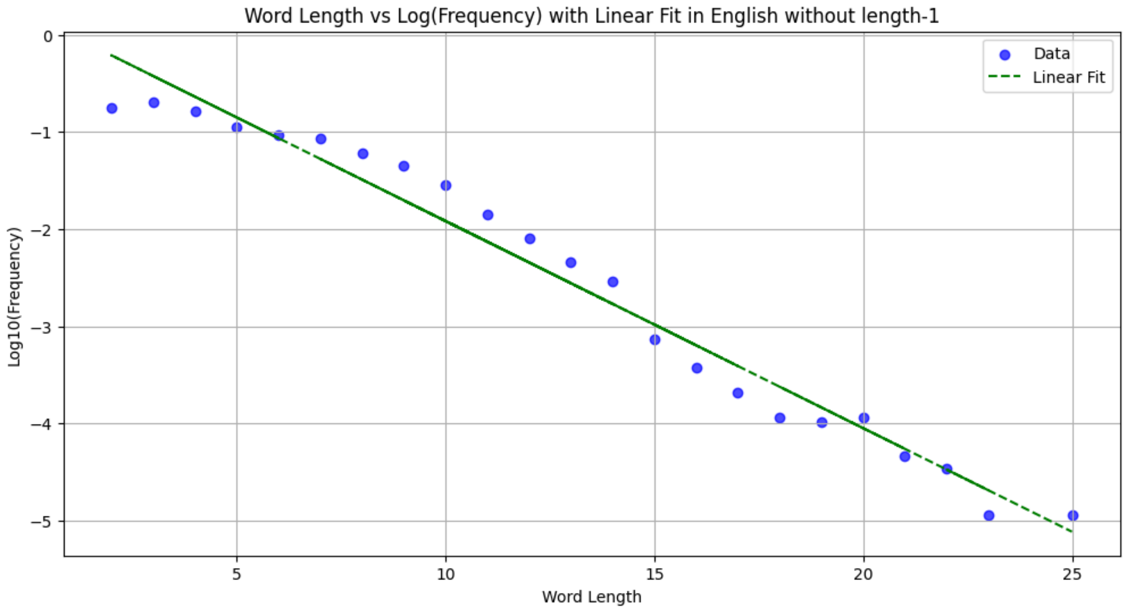

Figure 4: Linear regression plot for Scottish Gaelic, excluding words of length = 1

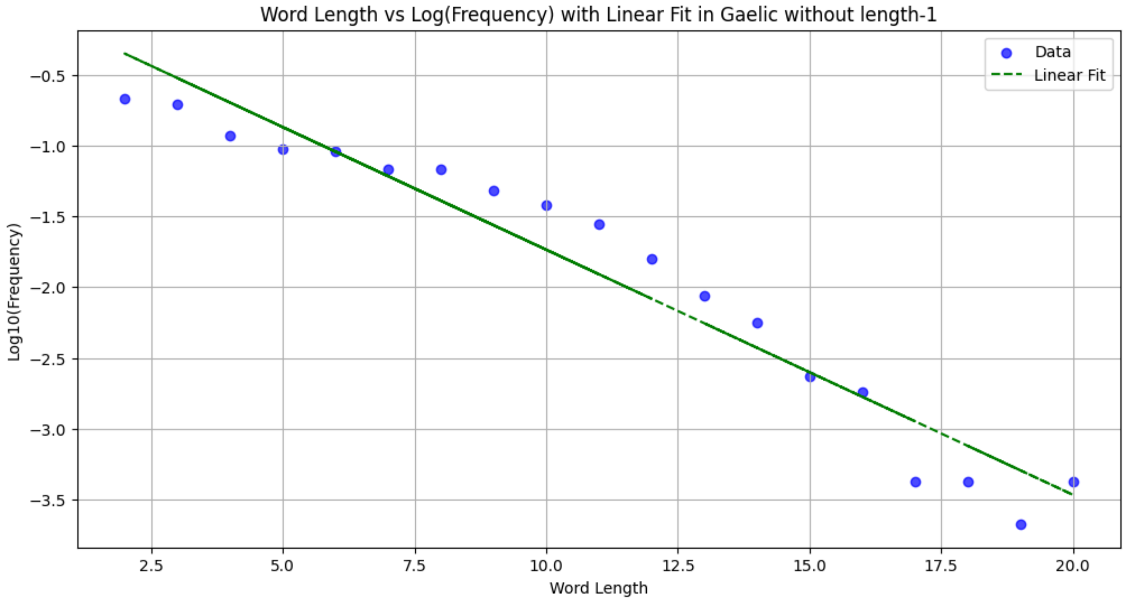

Table 1: Test results for all four samples:

Sample	Slope	R^2	KS 2-sample test
English	-0.2***	0.93	0.0741***
Gaelic	-0.16***	0.91	
English (no length-1)	-0.21***	0.97	0.0516***
Gaelic (no length-1)	-0.17***	0.94	

References

Francis, W. N. & Kučera, H. (1961). A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown).

Kanwal, J., Smith K., Culbertson J., Kirby S. (2017). *Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication.* Cognition, 165, 45-62. https://doi.org/10.1016/j.cognition.2017.05.001

Lamb, W., Arbuthnot, S., Naismith, S., Danso, S. (2020). *Annotated Reference Corpus of Scottish Gaelic* (ARCOSG), 1997-2020 [dataset]. University of Edinburgh. School of Literatures, Languages and Cultures. Celtic and Scottish Studies. Available at https://github.com/Gaelic-Algorithmic-Research-Group/ARCOSG

List of contributions

Coding / checking the writing: Eloi Singla i Milian

Writing / checking the code: Maria Leva Jiménez

Link to code and data repository

https://github.com/EloiSinglaMilian/NLP-Exercise-1/