Zipf's Law of Abbreviation in English and Chinese

Sijie Ju

1. Introduction

Zipf's Law of Abbreviation (ZLA) refers to a linguistic law that more frequently a word is used, shorter the word will be. The length of a word is proportionate to its frequency. I will test this law in English and Chinese. Different from alphabetic languages, Chinese is an ideographic language that characters are combined to form a word. Generally, the length of a word varies from 1 to 4. The number of two-character words are the largest. Therefore, I hypothesized that there would be some difference in result of these two languages. The text I will test on are news reports. Generally, the length of a news report is either too short or too long. ZLA is a statistic regularity, so if the text is too short, the absolute frequency will lack the statistical significance and be unable to support or oppose the law.

My hypothesis is that English will conform ZLA, but Chinese won't.

2. Material and methods

To test ZLA in English and Mandarin, the first step is to obtain two news report in both languages. I found two reports of similar numbers of words on the websites of two famous media, BBC for English and The Paper for Chinese. Secondly, I used the library *bs4* to scrape information from web pages. As this tool scraps all the text on the website, I cleaned it using regular expression. For the English text, I lemmatized it using spacy. Lemmatization was done instead of tokenization because tokens in different inflections are actually the same word and should be counted equally in this case. For the Chinese text, jieba was used to do the tokenization. As jieba is only a tokenization tool and is unable to recognize whether a token is a punctuation, I filtered them out with a list of punctuation. After obtaining the processed tokens, I counted the frequency and length of each token, organized the data into a table and finally visualized the data into a scatter plot to see the result.

3. Results

As is shown in Table 1, there indeed exists a negative correlation (p value are near zero in all cases) between the frequency and the length of a word in both languages.

Language	ρ
English	-0.236
Chinese	-0.160

Table 1 Pearson correlation coefficient in English and Chinese

Figures 1 and 2 illustrate the frequency distribution of words based on length in English and Chinese. In English, except the one-letter word: *I* and *a*, most of the words show the tendency that the more frequently used word will be shorter in length. The highest frequency of each length reduces as the length increases. Some exceptions occur, especially the word with 6 and 7 letters, but this is mainly due to the topic word (*cancer*) and the name of the main character.

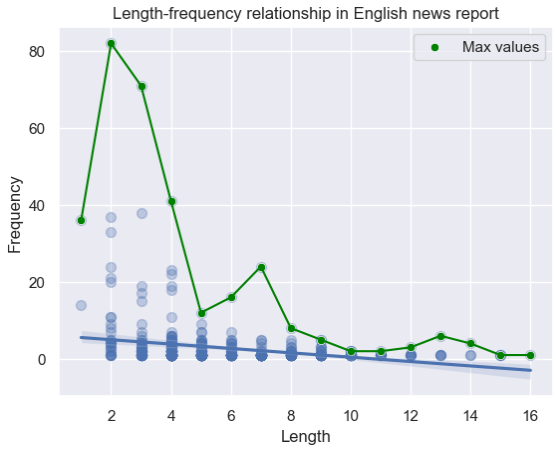

Figure 1 Length-frequency relationship in English news report

The case in Chinese is quite similar to English, indicating a decreasing frequency trend, which is beyond my expectation. After checking the length-frequency table, I observed that the frequently used one-character words are the grammatical particles and prepositions, which are crucial for sentence formation. However, despite Chinese also conforming ZLA, the frequency of two-character words is comparable to that of one-character words, and even higher within the range of 10 to 20.

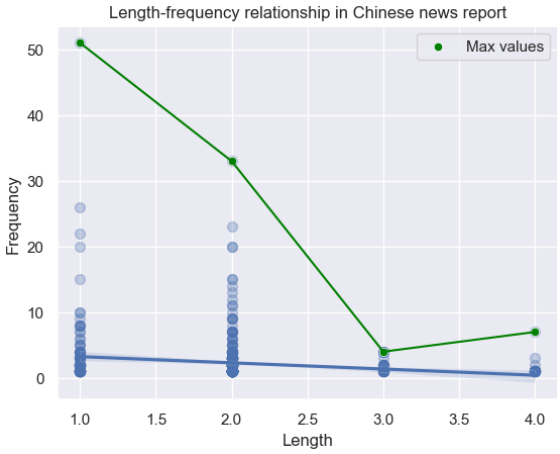

Figure 2 Length-frequency relationship in Chinese news report

Comparing the result in Chinese and in English, the negative correlation is more obvious in English. This can be partly attributed to the shorter length of Chinese words. The frequently used words are predominantly concentrated among those with one or two characters.

From my perspective, this result can also apply to other news reports. News reports should be clear, readable and transmit enough information to the readers in an appropriate length. Thus, easy and short words will be used more frequently in news reports, which conforms to ZLA.

This test still has some problems. First, the data scraped directly from media websites are still mixed with unexpected text, e.g. description of photos, recommended reports, which is impossible to filter them out without manual intervention. Additionally, the tokenization tools in both languages are not powerful enough to accurately tokenize all the words, particularly in Chinese. The absence of word spacing in Chinese leads to a great challenge for tokenization.

References

Fernández, A.H., Casas, B., Ferrer-i-Cancho, R., Baixeries, J., (2016). Testing the robustness of Laws of Polysemy and Brevity versus frequency. https://irdta.eu/slsp2016/Download/slides/TESTING%20THE%20ROBUSTNESS%20OF%20LAWS%20OF%20POLYSEMY%20AND%20BREVITY%20VERSUS%20FREQUENCY.pdf

Wikipedia contributors. (2023). Brevity law. *Wikipedia, The Free Encyclopedia.* Retrieved 19:13, February 3, 2024, from https://en.wikipedia.org/w/index.php?title=Brevity_law&oldid=1168576394