Alba Prados, Daniel J Casas, Yutong Zhang

Zipf's Law of Abbreviation

Introduction

With this project, we aim to analyze whether Zipf's Law of Abbreviation holds for two typologically different languages: English and Finnish. English is an isolating language, whereas Finnish is agglutinative.

Zipf's Law of Abbreviation is one of the statistical patterns observed in human languages, named after the linguist George Kingsley Zipf who popularized its study. As a universal structural property of language, this law states that the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences (Zipf, 1935; p. 23). In other words, more frequent words tend to be shorter and conversely, less frequent words are longer in length.

To test this law, we are conducting an experiment in which we will tokenize two texts, one for each language, and count the word frequency. We have chosen religious text as they have similar semantic fields, recurring themes and concepts.

We expect that Finnish will present a higher number of individual words due to the strategies of word formation they use, based on high morphological inflexions and particles; features not found in English.

Methods

The text in Finnish is called "Suomalaisten runojen uskonto" (The religion of Finnish poems) by Kaarle Krohn and the text in English is "Curiosities of Superstition, and Sketches of Some Unrevealed Religions" by W. H. Davenport Adams. Both were sourced from Project Gutenberg.

To process the raw .txt files, we used a series of libraries, like urllib, spacy, pandas, Counter matplotlib, and seaborn, adapting extra code to filter the unrelated introduction and copyright content from the respective text. These libraries allowed us to access the text, tokenize it, tabulate the results, and visualize the findings. To tokenize, we used the small pretrained pipelines for English and Finnish. Lastly, we used .strip () to get rid of unwanted spaces appearing as "\r\n". For Finnish we also performed additional cleaning to get rid of list items and intermittent English stop words with high frequency that we found in the corpus, like "the" or "of".

Let us describe now the treatment of the tokenized data. First, we took the list of common words in each language, then we created a DataFrame, calculated the length of each word, sorted the DataFrame by frequency in descending order, and added a rank to each row based on the sorted order. Let's have a look at the code used for English:

```
df_en = pd.DataFrame(common_words_en, columns=['word', 'frequency'])

df_en['length'] = df_en['word'].apply(len)

df_en = df_en.sort_values(by=['frequency'], ascending=False)
df_en['rank'] = list(range(1, len(df_en) + 1))
df_en
```

After the results were tabulated, we created a relational plot, where the x-axis represented the rank of words, and the y-axis represented their frequencies from the DataFrame. This visualization helped us understand how word frequencies were distributed across different ranks in the dataset.

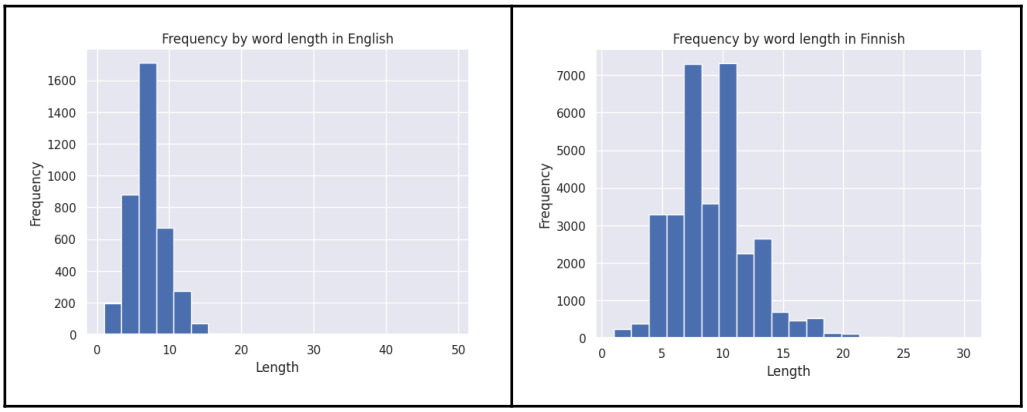
```
sns.relplot(x="rank", y="frequency", data=df_en);
plt.show()
plt.close()
```

Finally, we added one last column 'logfreq' to the DataFrame, containing the natural logarithm of the word frequencies with the addition of 1 to each frequency. This transformation was done to compress the scale of the frequencies and mitigate the impact of extreme values.

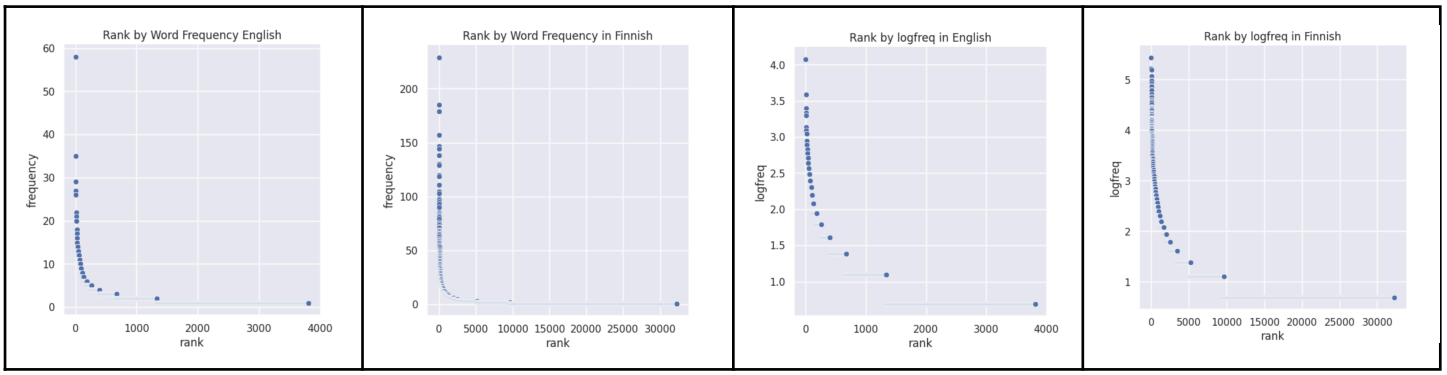
```
sns.relplot(x="rank", y="logfreq", data=df_en);
plt.show()
plt.close()
```

Results

Both English and Finnish texts indicate that Zipf's Law of Abbreviation holds successfully, showing the universal structural property of language.



According to the word frequency histograms, English shows a clear peak in word length at 5-7 characters, followed by a gradual decrease in the peak as word length increases. The result for Finnish, on the other hand, shows a particular pattern where the most frequent word length hovers between 7-8 and 10-11 characters. This aligns with the expectation that the morphological complexity of Finnish, as an agglutinative language, would produce more word forms due to inflection and compounding. Moreover, considering the unavoidable terminology with high frequency in religious texts might also affect the length of word, such as: “Buddha” and “Hiouen” in the English text; “Neitsyt”, “Jumalan” and “Väinämöisen” in the Finnish. Therefore, although the Finnish result is higher than the English peak, both trends confirm the essence of Zipf's Law of Abbreviations in terms of the overall distribution, with shorter words indeed occurring more frequently.



Furthermore, by analyzing the above two sets of plots concerning frequency versus rank and log-frequency versus rank, respectively, it can be seen that the former set, in both English and Finnish, shows a typical Zipfian distribution, where the most frequent words have the lowest rank and their frequency decreases sharply as rank increases. Meanwhile, the latter group, following the similar trend, reflects the principle that a small number of words are used very frequently, and many words are used infrequently.

Based on the above analysis, our results are largely consistent with Zipf's Law of Abbreviation, which leads us to believe that the results can be generalized to other similar data. However, we are also conscious of the limitations of our data and approach. First, due to language limitations, we were unable to find religious data with a similarly wide range of content in both languages; second, we failed to completely filter out religiously irrelevant content due to the confusing layout of the free e-books, a problem that occurs mainly in the Finnish dataset, with a small amount of English words inevitably mixed in. Finally, we filtered only the high frequency of low information content in both languages.

In conclusion, the patterns observed in the visualization are qualitatively similar, suggesting that Zipf's Law Abbreviation applies to both English and Finnish.

Github: https://github.com/00alba00/Exercise1_NLP

References

- Adams, W. H. D. (1828-1891). Curiosities of Superstition, and Sketches of Some Unrevealed Religions (No. 17016017). Project Gutenberg. https://www.gutenberg.org/ebooks/41566
- Krohn, K. (1863-1933). Suomalaisten runojen uskonto (No. 49028). Project Gutenberg. https://www.gutenberg.org/ebooks/49028
- Zipf, G. K. (1935). The Psycho-biology of Language, Vol. IX. Houghton Mifflin.

List of Contributions

Preparation:
- Alba: 33.3%
- Daniel: 33.3%
- Yutong: 33.3%

Report:
- Alba: 33.3%
- Daniel: 33.3%
- Yutong: 33.3%

Code:
- Alba: 33.3%
- Daniel: 33.3%
- Yutong: 33.3%

Posting Code to Github:
- Alba: 33.3%
- Daniel: 33.3%
- Yutong: 33.3%