Exercise One: Zipf's Law of Abbreviation

1 Introduction

The american linguist and philologist George Kingsley Zipf studied the relationship between a word's length and its frequency: he observed that the more frequent a word is, the shorter its length tends to be. He also claimed that this Law is supported by the Principle of Least Effort, which states that individuals are keen in optimising form-meaning mappings in order to communicate efficiently and accurately at the same time. He affirmed that the effects of such Law can be observed in a wide range of human languages, in computer programming languages and even in some systems of animal communication.
In this exercise I will try to test if Zipf's Law od Abbreviation will hold on a different type of text from the nove: I will test it on a poem, firstly on one in English and then one in Italian. I chose the poem because it has a particular structure and it can contain more rare and unusual (and possibly longer) words compared to a novel, so this can be a challenge for Zipf's Law. As the other language I decided to test Italian because it'is my native language, so I will be able to see how this Law holds on a language that I know so well and that has different linguistic properties from English. Also, it has a distinct length for most frequent words such as articles and conjunctions (i.e. il, la, lo, e, poi compared to the, and, then).
For the results, I expect that Zipf's Law will hold on the genre poem, since it has been broadly proven, but maybe with some discrepancy, especially in the italian poem, where I expect longer words.

2 Material and methods

The first thing I had to do was to get the data that I will analyze in order to test the accuracy of the Zipf's Law. I pick out two really famous and pretty long poems: I chose "The Raven" by Edgar Allan Poe and "Canto Notturno" by Giacomo Leopardi. In order to later work on the text of the poem I wanted to find it in the .txt format. For the poem "The Raven" I was able to find it on the Gutenberg Project website: since in the text contained also a commentary of the poem I just copied and pasted the text of the poem in a .txt file on VScode. For the poem "Canto Notturno" I could download a .txt file of the poem by *"it.wikisource.org"* and then I proceeded to delete the commentary in the end of the file.
I wanted to compare the frequency of each word in the text with its length (e.g. the number of letter of each word) to test Zipf's Law. Then I would compare the results for each poem. To start I had to tokenize the text: I firstly got rid of all the special characters, numbers and uppercase letters using regulars expression (I substitute everything that wasn't a letter with a space). To actually tokenize the text I used the library "spacy" and the language model specific for English and Italian. Then, I created the variable "doc", which contains the spacy natural language process of the poem text previously modified with RE. Finally, I associated the variable "tokens_en" or "tokens_it" to all the tokens found in "doc" under the condition that they weren't punctuation

signs nor spaces. In this way I obtained the tokens of the text representing just the (lowercased) words.
Afterwards, I build a dictionary with the length of the words as keys and their frequency as values so that I could later plot a bar diagram and actually see how the frequency of the words varied based on their length, thus test Zipf's Law of Abbreviation.

3 Results

Looking at the graphs (fig 1 and fig 2) we can draw some conclusion about the holding of Zipf's Law of Abbreviation, from which we expect a decrease of the frequency corresponding to the increase of the words length. Firstly I will analyze the the data of figure 1: we can see immediately that there is a pick of frequency (262) corresponding to the length "4" compared to the second (173) and third (163) most frequent length of words. This a significant distance from shorter words, especially from the shortest ones (length of 1 character), that are present just 58 times. For what concerns longer words, from length 5 on we can see a linear decrease of their frequency (length 6 and 7 differ just from 1 point of frequency).
If we proceed in the analysis of the data of the italian poem we can see a pretty different graph. A similar pick to the one seen in the English text plot can be observed in the length 2 and 5. But, differently from the previous pick these ones don't have such a big distance from the other most frequent length. Comparing the graphs of the two languages we can see that both of them don't show a linear decrease: the shortest words are significantly less frequent than the immediate following ones (in the english text more than in the italian one). Another similarity is that from the length 5 in the english text and from the length 6 in the italian one they both seem to hold very well the Law, with decrease of the frequency proportionally inverted to the length of the words.
Since the Law just tells us that the shorter words will be more common, and not that the shortest has to be the most frequent of all, we can conclude that the law holds in both poems, but we also have to underline that it holds at a significantly higher degree to the italian poem compared to the english one.
The reasons for such results can be various. Firstly we have to mention that poems are short text, so even though I tried to choose some that were pretty long, it's a smaller set of data compared to a novel, thus the results obtained could be less reliable. Moreover, the english poem (1102 tokens) is longer that the italian one (827 tokens), so this might result in the difference of their graphs. Another reason for such difference can be the use of diverse common words such as demonstrative in english, which typically have 4 letters (this, that), or articles in italian (il, lo, la) which have 2 letters.
Overall, my initial expectations about the results were met, even thought I would have expected a slightly bigger presence of longer words in italian. To conclude, I think that this results would generalize to other data of same genre of text: even with this small example we could see a big similarity in the results of these two poems. I'm not sure about their generalization to other languages: I think that in that case the results would differ more, especially between distant languages. Here I examined english and italian, and even if one of them in a romance language

and the other is a germanic, they have similar properties and words, but even in this case we could still see some substantial differences. Probably with languages such as russian, turkish, etc. the results in this kind of text wouldn't be the same.
Surely the outcomes here discussed and the methods used have some limitations: this test should be done with a lot more poems, with different length and style (also the period in which they were written influences the words used and their length). Probably other more complex statistical methods other than a diagram bar plot would be better to analyze the results. With this simple data and test I was able to draw some intriguing conclusions and observation, but this research in such an interesting topic can be extended largely.

Figures

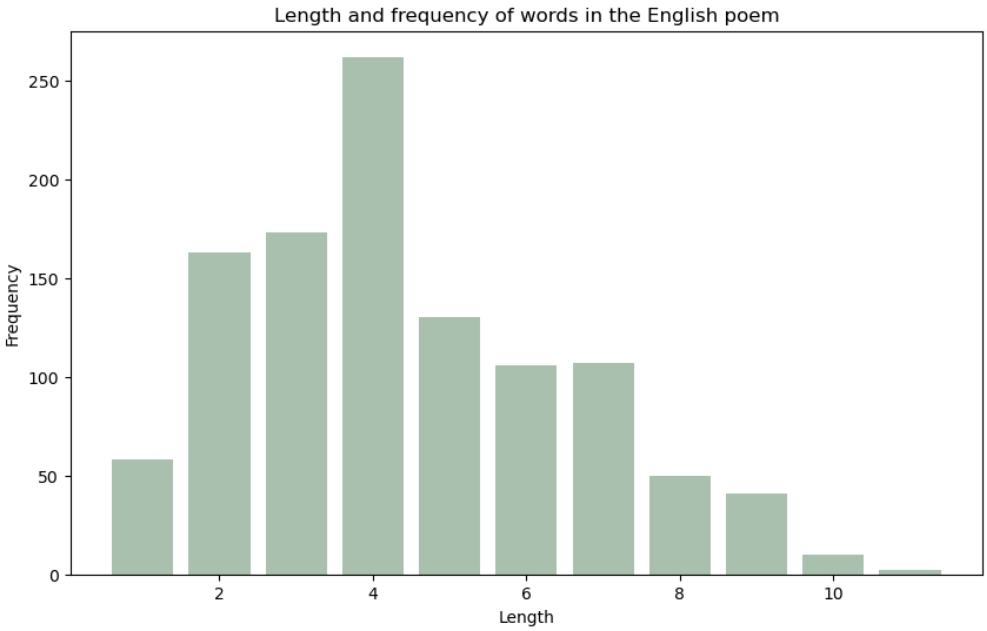

Figure 1: Plot of words in the english poem

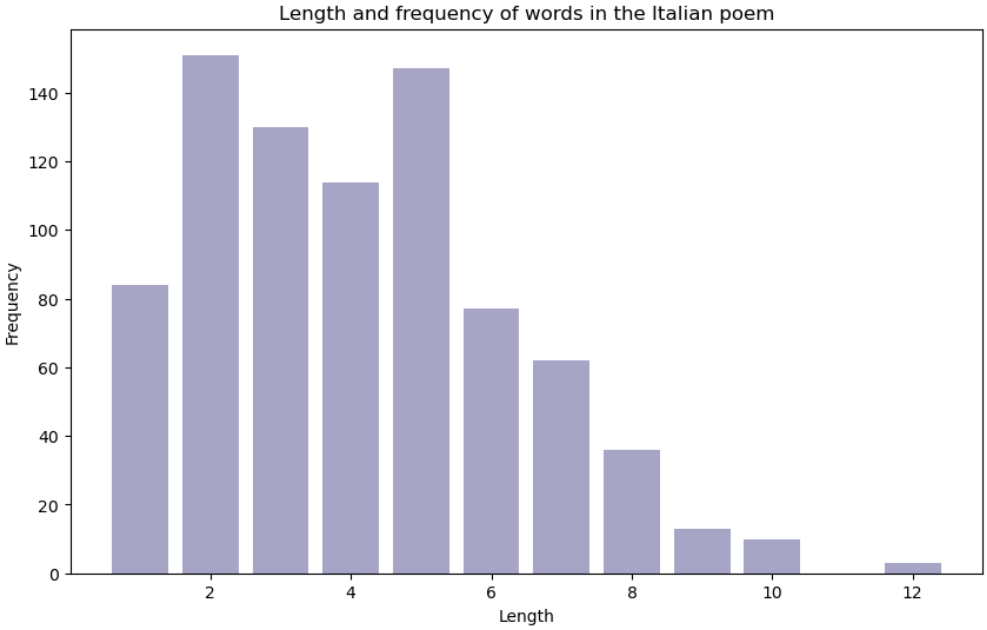

Figure 2: Plot of words in the italian poem