Application of Zipf Law on Turkish and English Language Versions of Harry Potter and the Philosopher's Stone

Nisanur Bulut *

*Computer Engineering/Bursa Uludag University, Turkey

*nisanurbulutnb@gmail.com

Abstract **–** This study aims to show the results of the zipf law on the written text in Turkish and English languages. According to Zipf's law, in a sufficiently long text written in a natural language, there is a special relationship between words, both in meaning and frequency of use. This relationship can be expressed with mathematical formulas. When we count the number of times the words in the text are used and then sort these words in descending order of frequency, we divide the frequency number by the ordinal number and we always get a fixed number. Zipf's law consists of 4 principles. These are: fixed number representation (f*r=c), frequency ratio representation, probability distribution representation (f/r) and frequency indicator based on word meanings. In this study, the first three principles of the law were used. The book Harry Potter and the Philosopher's Stone was chosen to enforce this law. Both formulas of Zipf's law have been applied separately to the English and Turkish versions of this book. In the study, the law was applied by making both word and suffix analysis and the results were shown with graphics. A desktop application was developed in .NET technology for demonstration. The Zemberek library was used to examine the written texts and to determine whether the words were Turkish. In addition, meaningless conjunctions and proper nouns and titles were not included in the study.

Keywords - Zipf, natural language processing, zemberek, frequency

I. INTRODUCTION

In the field of natural language processing, it has been noticed that there is a relationship between elements such as the frequency of use of words, the number of meanings a word carries, and the distances between words. The first studies in this sense were made by George K. Zipf (1902–1950). Zipf published these studies in his article titled "Selected Studies of the Principle of Relative Frequency in Language" published in 1932 and became law under his name.[1] According to Zipf, when we sort the words in a text in descending order of frequency of use, regardless of the content and language, a list with a certain pattern is obtained. The order in which a word is used in the list multiplied by the frequency gives a fixed number. This law was first applied in James Joyce's Ulysses, as follows; while the most used word is used 2653 times in the text most used 265 times most used, used 133 times.

According to Zipf's findings, while the majority of words are used very rarely, a small number of words are always used.

II. MATERIALS AND METHOD

Turkish and English versions of Harry Potter and the Philosopher's Stone are the main materials of the study. To apply the Zipf law, a desktop application was developed in C# programming language and the test results were shown with graphics. The source codes of the application are at https://github.com/NisanurBulut/ZipfLaw_NLP github.

A. Methods Of Zipf Law

Unique words in a text, that is, the number of times a word occurs in the text is counted, provided that it is not counted more than once. Freq(word)=Indicates how many times a word is used in the text.

According to this frequency list, each word is ranked in descending order. While the rank of the most used word is 1, the rank of the most used word after it is 2. We are not concerned with the meaning of the word, only with the frequency and degree values.

r=rank(rank)

N = the number of words in the text. It is not a unique word count, so a word can occur more than once in the text and is included in the count.

Probe(r) = freq(r) / N According to Zipf's law;

r * Probe(r) = A

According to this formula, A is a constant number and usually takes the value 0.01. Zipf's law is a statistical inference, it always gives the same value.

It considers the formula probe(r)=freq(r)/N and
If we rewrite Zipf's law,

The formula r * freq(r) = A * N is obtained.

If we apply Zipf's law, the first thing we should get is the freq(r) value. When we multiply this value by degrees, r, the results we will obtain are approximately the same. This does not mean that the r*freq(r) value for each word will be the same. But the results are expected to be close to each other.

It is tricky to look at the results for the words with the most or the least frequency value. The results for these are usually the ones with the most errors. For ideal observation, the degree of words in the order of 1,10,100,200… should be checked for their conformity to Zipf's law. What suitable notation for Zipf's law are point graphs. It will be seen more clearly when the results are displayed after taking the logarithm.[3]

In the application phase, the principles of Zipf's law, which have been explained above, have been tried to be shown.

B. Application Program on Zipf Law

The program tries to show the English original of HarryPotter and the philosopher's stone and the compatibility of the word-roots in the Turkish translation with Zipf's law. In addition, for the Turkish translation of the book, it lists the affixes of each word with its types and shows its compliance with the zipf law. Pre-Operations on the Text

Book texts in pdf format were converted into txt files.

All words in text files have been converted to lowercase.

Removed punctuation between words.

While researching Zipf's law, words were separated into their roots and a calculation was made over their roots.

The words were checked whether they belong to the Turkish language with the help of a spring. For this reason, words such as Harry, Privet, Mr, Mss are not included in the count. Thus, it is exempted from a transaction burden such as checking their personal names.

C. Components Used in the Development of the Program

The .Net format of the Zemberek library was used to make word analyzes on the written text. LiveChart Framework was used for the graphical representation of the results. The iTextSharp.Net library was used to store the results of the research on the compliance of the Zipf law in pdf file format.

The program has been developed as a desktop Windows Form Application using the C # programming language.

In the result displays, the distribution of the first 30 words is shown.

*D. Representation of f*r=A on Turkish Text*

In the Turkish text, each word was counted and ranked in ascending order according to the frequency of use. Multiplying the amount of use by its degree should give approximate numbers. Circle graphic representation was used to make it clearer that the results give the same values in a close sense.

In the Turkish text, 55449 non-singular words were found and the number of singular words was determined as 13088. When the root distinction is made on these words, the number of unique roots is determined as 2340. Roots are classified by their types such as adjective-root, noun-root, preposition-root, pronoun-root, conjunction-root, tense-root, question-root, exclamation-root, spur-root, echo-root, number-root, etc. counted.

III. DISCUSSION

This should explore the significance of the results of the work, not repeat them. The results should be drawn together, compared with prior work and/or theory and interpreted to present a clear step forward in scientific understanding. Combined Results and Discussion sections comprising a list of results and individual interpretations in isolation are particularly discouraged.

In the Turkish text, the words are divided into their roots. And the number of each root depending on its type is shown in figure 2. Since root research was not done in the English text, no listing based on root type was made.

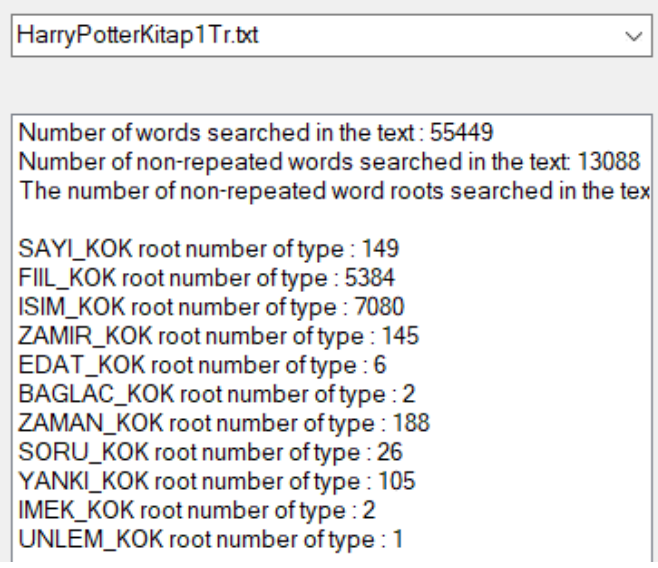

Fig. 2 Number of root type in Turkish text

After counting the root word roots in the Turkish text, the ordering for the most used root is rated starting from 1. The fixed number obtained as a result of multiplying the frequency of use by its degree is definitely not the same for the first 30 words, but as it can be seen from the graph, it has close values.

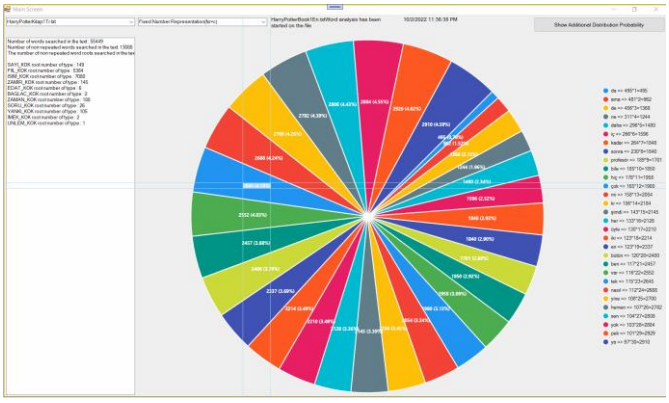

Fig. 3 Number of root type in Turkish text

E. r(f/N) Representation on Turkish Text*

In the Zipf law explanation section, it was said that when the formula is rewritten, the formula r * freq(r) = A * N is obtained. In the implementation of this rewriting by the program, it has been tried to show that there is a constant A number again indirectly. The decimal values found as a result of multiplying the frequency of the word roots to the number of single words in the text were rounded to 4 digits. Although it is not expected to show the exact same value, it is aimed to find close values.

When the figure 4 is carefully examined, it is seen that the decimal values found are close to each other, although they are not exactly equal to each other.

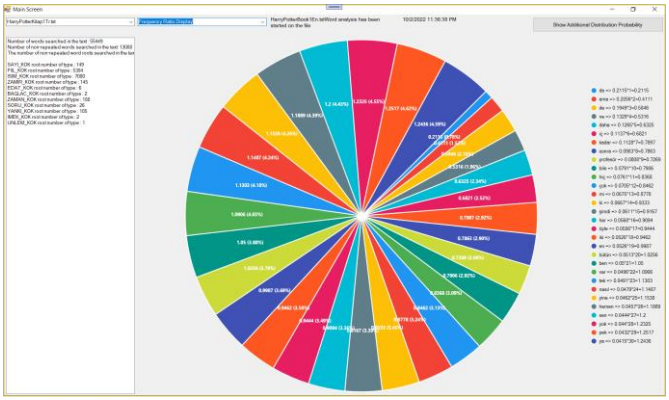

Fig.4 Number of root type in Turkish text

F. f/r Representation on Turkish Text

When we rank the word roots according to their frequency of use and divide the number of frequencies by the degree, it is expected that the values obtained will form a folded pattern. It is

expected that the results will have close values even if they are not at the exact same fold values.

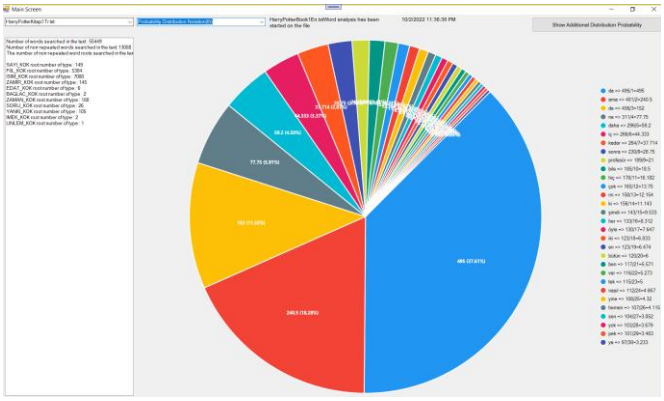

Fig. 5 Number of root type in Turkish text

When the figure 5 is examined, the root "da" in the first row has a ratio value of 495, while the "but" in the second row has a ratio value of 280.5. The “de” in the third row is worth 152, and the “ne” in the fourth row is worth 77.75.

G. f/r Representation of Word Suffix Distributions on Turkish Text

Each word is divided into its suffixes with the help of Zemberek library. During the annealing process, each affix was counted with its type. Considering the distribution of affixes on the Turkish text, 24655 affixes were counted. When these suffixes were counted again without repetition, it was seen that there were 97. Looking at the types of affixes, 97 species were found again.

Attachments are listed on the left side of the screen, depending on their type. The circle graph in figure 6 shows the zipf law compliance of the attachments. The first-order supplement is almost double the second-order supplement, while the third-order supplement is three times as much. Even though the ordinal frequency pattern could not be represented with exact numbers, close values were still reached. On the far right side of the screen, there is the label display of the circle graphic. The f/r formula used for the acquisition of the first 30 values forming the graph is also dynamically written.

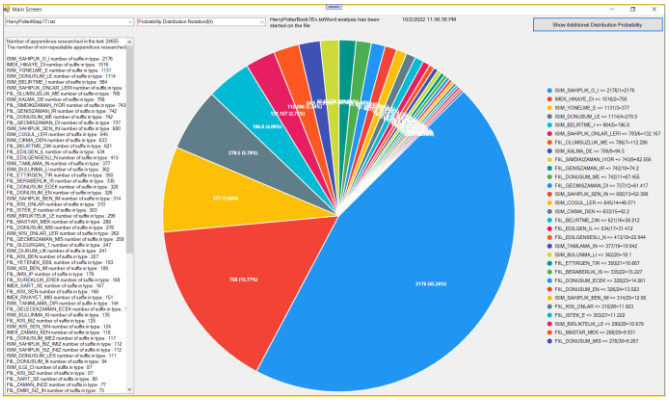

Fig. 6 Number of root type in Turkish text

H. f/r Representation on English Text

While researching the zipf law on the English text, the same pre-processes were applied in the Turkish text. However, of course, no English control was made on the words or no affixes or root divisions were made. Words are counted and reordered. The decimal values obtained by dividing the frequency numbers by their degrees are again rounded to four digits.

When the results found were evaluated, a definite folded pattern was not obtained, but considering the circular graph, the first word was approximately twice the second word, while the third word was three times as much. The same pattern applies to words in other rankings as well.

The graph in figure 7 representation of Zipf's law on the order frequency distribution on the English text is as follows:

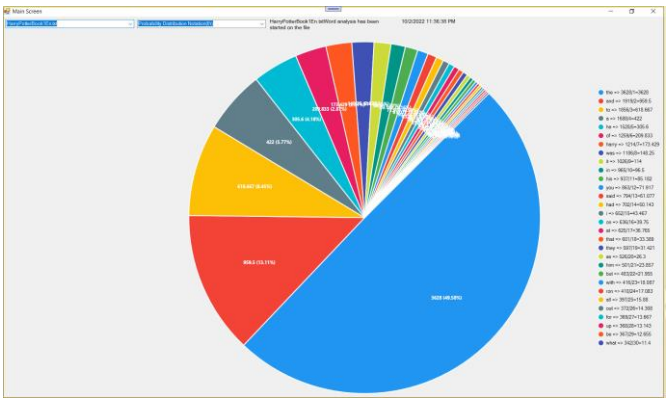

Fig. 7 Number of root type in Turkish text

IV. CONCLUSION

When we look at the ecosystem, there are many laws that have stability in themselves, for example the golden ratio, pi number, Euler's constant. These

laws are discovered laws, not laws laid down in an order. There is a strange order in nature that we have difficulty in comprehending in our lives. High intelligences continue to run after these secrets with inquisitiveness.

It should not be difficult to perceive the relevance of the results of the application of Zipf's law to cities or the existence of the golden ratio, because it speaks of a natural process, a piece of life belonging to living things. However, when we look at the text analysis, the mystery of why the fixed results produced by the law are obtained, regardless of language and content, is still preserved.

The selection tendencies hidden in our genes can be shown as the reason for this order frequency rule. But here, too, a stalemate is reached, because why and how these electoral trends also acted is still a mystery. Perhaps the zipf law is not a cause and effect, but rather an aspect that shows the result of our social behavior, communications, and business matters.

It is a fact, however, that solving the working mechanism of zipf's law will predict much more than what human tendencies are.

“The laws of nature are written in the language of mathematics by the hand of God.” “To understand nature, you must understand the language in which it was written, and that language is mathematics.”

Galileo Galilei

REFERENCES

[1] *Leitch, Matthew (2010), A Pocket Guide to Risk Mathematics: Key Concepts Every Auditor Should Know, John Wiley & Sons, p. 62, ISBN 9780470971468*.

[2] Zipf Dies After 3 - Month Illness, The Harvard Crimson, September 27, 1950

[3] Saichev, A. I.; Malevergne, Yannick; Sornette, Didier (2009), Theory of Zipf's Law and Beyond, Lecture notes in economics and mathematical systems, vol. 632, Springer, p. 1, ISBN 9783642029462.