Lecture Notes on Language and Literature (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2023.061009
ISSN 2523-5869 Vol. 6 Num. 10

The Applicability of Zipf's Law in Report Text

Zhang Yang*, Zhu Xiangyi

School of International Studies, Zhejiang University, Hangzhou, China
**Corresponding author*

Keywords: Zipf's Law, report text, AntConc, report of the 20th National Congress

Abstract: Zipf's Law, discovered by Harvard linguist Zipf in the 1940s, is an empirical Law reflecting the general rule of word frequency distribution in language. It describes the link between word frequency (F) and its frequency rank (R) as F×R=C (constant)[1]. A large number of studies have been conducted to test the applicability of Zipf's Law in various languages and diverse text types. However, as language keeps developing and changing, there is a continuous need for the verification of Zipf's Law in the most up-to-date materials and texts in different languages. In addition, different categories of texts should be taken into consideration. Based on this, this study selected the report of the 20th National Congress and its multilingual translations, which contain English, Spanish, and Russian, as our research objects. We used AntConc to detect the distribution of their word frequency and found that all of these report texts of four languages conform to Zipf's Law, and there is a slight difference in terms of fitness degree between the Russian text and the other three. We assumed that it is mainly caused by the limited text length and the translation process.

1. Introduction

In recent years, a large number of studies have confirmed the applicability of Zipf's Law in different languages and different text types. However, due to the continuous development of language, the characteristic of different language texts is correspondingly changing too, which results in diverse distribution of high-frequency words. Therefore, the verification of Zipf's Law must be ongoing along with the course of the evolution of language. To our knowledge, the majority of previous research mainly focused on the analysis of literary text and spoken text[2], while a relatively limited number of research has been done on report text, and more specifically there is a gap in the research on Chinese report text, which highly represents the latest formal language pattern of Chinese and reflects the most heated social issues and public concerns. Report text is characterized by its formal language and most up-to-date content. It can be seen as the newest product of language evolution, for it contains a large number of new words and fresh expressions born in the latest social background. Based on the features mentioned above, report text is one of the most appropriate types of text to verify Zipf's Law's consistent applicability with the continuous development of language. With the current study, we filled the research gap in the applicability of Zipf's Law in Chinese report text and its translations. To be specific, we analyzed the report of the 20th National Congress and its multilingual translation versions, which contain English, Spanish, and Russian. We used AntConc and Altmann-Fitter to detect whether their lexical frequency distribution conforms to Zipf's Law, and we drew a comparison between the fitting

degree of those four texts. Furthermore, we tried to analyze the possible reasons for the difference in fitting degree based on the text length and text origin.

1.1 Zipf's Law

Zipf's Law was first discovered and illustrated by Dr. Zipf himself, who found out that the relationship between word frequency and word frequency rank in a large text follows a concise and simple mathematical expression, that is $f(r) \propto r^{-\alpha}$[3]. Later, Mandelbrot proposed and derived a generalization of this Law that more closely fits the frequency distribution in language by shifting the rank by an amount β[4], the mathematical expression, therefore, becomes: $f(r) \propto (r+\beta)^{-\alpha}$. The Zipfian Distribution is thus also referred to as Zipf-Mandelbrot Distribution. Through decades of study, it has been gradually proved that the Zipf-Mandelbrot Distribution is universally applicable in a large number of languages, with a slight difference in the value of α and β. The applicability of the Zipfian distribution is shared among almost all languages, and previous studies have also found that the actual word frequency curve also shares some commonalities. For instance, it has been found that all the word frequency curves of diverse languages have three segments, "with the lower segment invariably bending downward to deviate from theoretical expectation, and each segment demonstrating distinctive linguistic properties and different biases in use"[5].

Numerous linguists and scientists have conducted works to provide a thorough and convincing explanation for the underlying mechanism of Zipf's Law. The first attempt to explain the mechanism of such linguistic phenomena is held by Zipf himself, who considered the Zipfian distribution as the outcome of the negotiation between the speaker and the listener for optimal communication[6]. It is believed that under the "least effort principle", the speaker wants to use as few words as possible in communication, whereas the hearer wants to hear as many words as possible to grasp the maximum amount of information. As a result, an equilibrium is attained between the listener and the hearer, whose conversation now reflects a near-Zipfian distribution. A similar view is also proposed by Mandelbrot[4], who also believed that the least effort principle is the reason behind Zipf's Law. However, the least effort principle is not the only mathematic model that was proposed to explain Zipf's Law, as Yule[7] and Simon[8] also introduced a simple stochastic model to explain. Similar to Zipf's idea, the simple stochastic model introduced two conflicting forces dominating the word frequency rank in the text, which are (1) the probability of reusing the previous words, and (2) the probability of introducing a new word. The outcome of such a mathematical hypothesis fits roughly with the Zipfian Distribution. Other attempts of uncovering the mechanism of Zipf's Law include Corominas-Murtra and Sol é[9], who proposed an explanation based on algorithmic information theory[10]. Such an attempt of explaining the mechanism of Zipf's Law, which is derived completely based on mathematical hypothesis, though excellently done, still left some important questions unanswered. The major question is how are these mathematical hypotheses linked with human psychological and cognitive processes, without a convincing explanation of which the study will still be merely examining the empirical evidence.

Works are also done to verify the applicability of Zipf's Law not only in different languages but also in different types of texts. Li et al. have testified to the applicability of Zipf's Law on the prescription of Traditional Chinese Medicine[11], whereas Linders and Louwerse[12]have conducted a similar applicability test, particularly on the spoken dialogs, and Qin[13]examined the applicability of Zipf's Law in ancient Chinese text, using the renowned ancient Chinese classics "Meng Xi Bi Tan" as the corpus. In reviewing the previous studies, we have found out that the applicability of Zipf's Law on the text of a political report is seldomly mentioned, in particular, the applicability of Zipf's Law on different translations of the same political report text has not yet been done, which provides us with an opportunity to testify the universality of the linguistic theory on a different type

of text.

1.2 The report of the 20th National Congress

“Hold High the Great Banner of Socialism with Chinese Characteristics and Strive in Unity to Build a Modern Socialist Country in All Respects” is the theme of the report delivered in the 20th National Congress on 16 October 2022. In terms of content, this report covers a wide range of areas, including social systems, education, health care, culture, military and ecological problems, and environmental protection. Each topic is accompanied by detailed background information and a description of its latest status.

The language used in the report includes a large number of new words and expressions, such as "Lucid waters and lush mountains are invaluable assets". It would be valuable to explore whether the frequency distribution of these new words conforms to Zipf's Law. In addition, this report also contains many words and expressions with Chinese characteristics, such as "one country, two systems", "The belt and road", etc. These words represent China's corresponding policies and systems, which are highly characteristic of the current developing trend and national specifics. They are the products of the new era and the best symbol of how language changes with the times. Therefore, exploring the applicability of Zipf's Law to the frequency distribution of these words helps to verify the universality of Zipf's Law along the process of language's continuous development. From the perspective of language characteristics, the report of the 20th national congress uses accurate words, and highly concise language with a precise meaning, which makes it an excellent report text sample.

As mentioned above, to our knowledge, there are still few studies aiming at the applicability of Zipf's Law on Chinese political report texts, so the report of the 20th national congress is an appropriate choice of report text to be analyzed for the reason that it is the most up-to-date and representative one.

1.3 The present study

In this study, we analyze the word frequency distribution of the report of the 20th national congress and its multilingual translations: English, Spanish, and Russian. Specifically, we seek to address the following two research questions:

(1) Does the distribution of word frequency in these four report texts confirm to Zipf-Mandelbrot distribution?

(2) What's the possible reason for the difference between the fitness degree of different language texts?

2. Methods

To analyze the fitness of the 4 versions of the report of the 20th National Congress to the Zipf-Mandelbrot distribution, we first obtained the word frequencies and word frequency sequences of the reports of these four languages, and then input the word frequency sequences into the distribution fitting software to check whether they can fit the Zipf-Mandelbrot distribution.

To obtain the word frequency and word frequency sequences of the reports in the four languages, we used AntConc to analyze the text. It is worth noting that before analyzing the Chinese text, word frequency analysis could not be performed for Chinese without prior word separation, so the Chinese text had to be separated into words first, and then word frequency analysis could be performed. Therefore, we used CorpusWordParser, a word splitting tool provided by Corpus Research Group of Beijing Foreign Studies University (http://corpus.bfsu.edu.cn/TOOLS.htm), to

first split the Chinese text; after getting the results of word splitting and word annotation, we imported the txt document into the word document, and replaced all the word annotation markers with blanks, and finally get the Chinese text without word markers, instead with words already split.

The Altmann-Fitter is an interactive software for the iterative fitting of univariate discrete probability distributions to frequency data. The current version of this software has over two hundred different distributions, and the Zipf-Mandelbrot distribution is one of them. After clicking the fit button, the table at the right end of the main window of the software shows the actual and theoretical word frequencies of the word frequency sequences, and after clicking the graph button, you can observe the relative positions between the actual and theoretical word frequency curves, which is a convenient way to visually study the fit of the text to the Zipf-Mandelbrot distribution.

At the end of the automatic fitting, we compare the specific values of the goodness of fit, coefficient of variation, etc. of the fitting of different texts with the Zipf-Mandelbrot distribution, and also give some descriptions of the two important parameters of the Zipf-Mandelbrot distribution for different texts, to compare the small differences in word frequency sequences between texts of different languages.

3. Results

Overall, the word frequencies of the Report of the 20th Congress for all four languages roughly conform to the Zipf-Mandelbrot distribution, and their respective goodness-of-fit falls within the acceptable range. (The fit is good when $R2 > 0.90$, better when $R2 > 0.8$, acceptable when $R2 > 0.75$, and unacceptable when $R2 < 0.75$.) The goodness of fit for all four languages exceeds 0.8, with English, Chinese, and Spanish texts exceeding 0.9.

When we examine the fitness of these four languages to the Zipf-Mandelbrot distribution, we can find that there are still some differences in the goodness-of-fit of these four languages, and there are also some interesting similarities and differences in the relative positions between their word frequency curves and the fitted curves. First, the best fit was found for English, with R2 equal to 0.9842 and the smallest Coefficient of Discrepancy C of 0.0239. The second best fit was found for Spanish, with R2 equal to 0.9622 and the Coefficient of Discrepancy C of 0.0432. The third best fit was in Chinese with a goodness of fit of 0.9509 and The Coefficient of Discrepancy C of 0.0380. The worst fit was in Russian with a goodness of fit of 0.8475, which is still within the acceptable range, and The Coefficient of Discrepancy C of 0.0544.

Here are some specific details of the word frequency curves for each language text. First, the Chinese version. First, there are 3213 words in total in the report of the 20th National Congress, with R2 of 0.9509 and C of 0.0380. Regarding the two parameters of the Zipf-Mandelbrot distribution curve, a is 0.8822 and b is 1.8816, which means that the relationship between word frequency and word frequency sequence in the Chinese language should be: $f(r) \propto (r+1.8816)^{-0.8822}$. Second, by observing the actual word frequency curve and the relative position of the theoretical word frequency curve, we can find that in the word frequency sequence of Chinese text, from word 2 to word 36, the actual word frequency is smaller than the theoretical word frequency, in which the difference between the fourth word and the ideal word frequency is the largest (the actual word frequency is 166 times, and the ideal word frequency on the fitted curve is 264.93 times); while from word 37 to word 1085, the actual word frequency is larger than the theoretical word frequency; from the 1086th to the 3213th word, the actual word frequency is smaller than the theoretical word frequency. From the graphs of the fitted curves, it can be found that the shape of the fitted distribution curves and the actual word frequency curves bear a resemblance to a cross in the beginning section, and their distance gradually decreases until after the middle section.

Next, the English version. First, the English version of the report of the 20th National Congress

has a total of 3081 words, with R2 of 0.9842 and C of 0.0239. Regarding the two parameters of the Zipf-Mandelbrot distribution curve, a is 1.0232 and b is 0.9128, which means that the relationship between the word frequency and the word frequency sequence of the English language should be: $f(r) \propto (r + 0.9128)^{-1.0232}$. The actual word frequency of English fits splendidly with the theoretical word frequency, with the actual word frequency floating up and down tightly around the ideal word frequency curve, without any large deviations which are found in the word frequency analysis of Chinese.

The third is the Spanish version. The report of the 20th National Congress for Spanish has a total of 4283 words, with R2 of 0.9622 and C of 0.0432. Regarding the two parameters of the Zipf-mandelbrot distribution curve, a is 1.0323 and b is 0.2937, which means that the relationship between word frequency and word frequency sequence for the Spanish language should be: $f(r) \propto (r + 0.2937)^{-1.0323}$. The shape of the word frequency curve of Spanish and the fitted curve shows a similar scissor-like shape to that of Chinese, to be more specific, starting from the 5th word, the actual word frequency is larger than the theoretical word frequency, starting from the 11th word the actual word frequency is smaller than the theoretical word frequency, and starting from the 205th word, the actual word frequency is larger than the theoretical one.

Finally, there is the Russian version. The fit result for the Russian text is relatively less desirable but still acceptable. Specifically, the report of the 20th National Congressfor Russia has a total of 5818 words, with an R2 of 0.8475, and a C of 0.0544. Regarding the two parameters of the Zipf-Mandelbrot distribution curve, a is 0.8504 and b is 0.6152, which means that the relationship between word frequency and word frequency sequence for this language should be: $f(r) \propto (r + 0.6152)^{-0.8504}$. The actual word frequency curve and the fitted curve are similar to that of Spanish and Chinese, where there is a clear bifurcation. At first, the actual word frequencies are larger than the theoretical word frequencies, and then the actual word frequencies are smaller than the theoretical word frequencies. Since the goodness of fit of the Russian text is not as good as that of Chinese, Spanish, and English, the gap between the actual word frequencies and the theoretical word frequencies is larger for all words.

By comparing the actual word frequency curves and the shapes of the fitted curves for the four languages, we can find some obvious commonalities: the word frequency curves all fit poorly in the beginning section, best in the middle section, and a jagged line with plenty of right angles on the graph appears at the end of the curve. The observation that the word frequency curves fit best in the middle section is consistent with previous literature summaries[14] As for the jagged lines at the end of the curve, after examining the specific word frequencies, we found that they are due to the presence of multiple straight lines parallel to the horizontal axis, for the reason that there exists a large number of low-frequency words with the same word frequency stacked at the end of the word frequency curve.

Based on the results, it could be observed that the jagged lines at the end of the curve all appear and bear great resemblance to each other, for the reason that there exist plenty of words that have relatively low word frequency and have the same word frequency. Taking the Chinese text as an example, at the word frequency rank, from the 657th rank to the 822nd rank, all the words have the same word frequency, which is four times. From the 823rd rank to the 1085th rank, all the words have appeared five times in the Chinese text. These words with the same word frequency have formed different lengths of straight lines paralleled to the horizontal axis, and the same situation is also found in the English text, the Russian text, and the Spanish text. What is different is that the word frequency curve of the English text fluctuates tightly around the theoretical curve, without great fluctuation which has been found in the other three languages.

Since the mechanism of Zipf's Law is not yet clear, we can only speculate on the specific differences in the goodness-of-fit and the parameters of the fitted curves for different languages.

First, the difference in goodness-of-fit is presumed to be due to the small size of the corpus. In the corpus involved in the statistics, there are 41,873 words in Russian, about 32,000 characters in Chinese, 25,000 words in English, and 29,000 words in Spanish. In previous studies, the text size fitted with the Zipf-Mandelbrot distribution usually exceeded one million words, so the difference in the goodness-of-fit may result from the difference in corpus size. The second issue that awaits to be discussed is the difference in the two parameters α and β of the fitted curves for different languages. Previous studies have shown that for different languages, there are subtle differences in the α values when fitting to the Zipf-Mandelbrot distribution. For example, the statistical analysis of word frequencies for the European Union Charter for 21 languages, presented in a paper by Professor Feng, showed that these 21 languages differed in their α values. Among them, the α values of English and Spanish are similar to the results obtained in this paper. In the word frequency analysis of the Charter of the European Union, the alpha value for English is 1.11 and for Spanish is 1.04, while in the analysis of the word frequency of the report of the 20th National Congress, the α value for English is 1.02 and for Spanish is 1.03.

4. Discussion

This study mainly explored the applicability of Zipf's Law in the report of the 20th National Congress and its multilingual translations. As mentioned above, among the existing research on Zipf's Law, the number of studies conducted on the applicability in report texts is still relatively limited, so in this aspect, our study filled this gap well. In addition, this study selected the report of the 20th National Congress as the analysis object, mainly because it is the most up-to-date and representative one among all the reports that are currently available, for the reason that it has the best language quality, and at the same time perfectly reflects the newest characteristics of language born in the continuous development trend. As for the target language of the translation text, we selected English, Russian, and Spanish, for these three languages all have a large number of speakers, thus can encompass the characteristics of different linguistic families and groups and are therefore representative and relevant for analysis. As for the research methods, this study used AntConc to analyze and count the word frequency of the selected texts to explore whether their distribution conforms to Zipf s Law, aiming to answer the following two research questions: (1): Does the distribution of word frequency in these four report texts confirm to Zipf-Mandelbrot distribution? (2): what's the possible reason for the difference between the fitness degree of different language texts? We responded to these two questions based on the results of the data obtained by AntConc. Firstly, for the first research question, we found that the distribution of word frequency in the report texts of all four languages conforms to Zipf's Law, and the specific fitness degree data obtained by AntConc is 0.9509 for Chinese, 0.9842 for English, 0.9622 for Spanish and 0.8475 for Russian. This leads to our second research question: why there is a difference in fitness degrees and what are the possible reasons for that? Based on the size and origin of the report text, we give two possible explanations: first, the length of the selected report texts is relatively limited. According to the character count, the total number of characters in the Russian version is 21,502, which is a quite small number from the perspective of corpus research, and therefore the analysis results are more likely to be contingent. Because of the limitations of the length of the corpus, it may happen that the report texts selected are not representative enough for they cannot cover all the characteristics of the target language. This thus leads to a relatively lower fitness degree. For this question, further verification can be done in subsequent studies by increasing the number of report texts. Second, the possible reason for the difference in fitness degree of diverse languages may also result from the fact that the English, Spanish, and Russian report texts are translated from Chinese, rather than written directly in the target language. The translator of the report may not be a native

speaker of the language, so there may remain lots of translation traces in the final report text. What's more, the report itself aims at describing the heating issues in Chinese society and answering the most concerned social topics proposed by Chinese people, so when it comes to the report content, it relates closely to the situation and status quo of China, rather than Russia. Based on this, even though the translated report is one hundred percent authentic and grammatically perfect, it cannot be viewed as a perfect sample of the Russian language, for it doesn't contain any up-to-date social issues discussed by the Russian people. As a consequence, we can see that the translation text may not be that qualified to represent the target language. This concern may give another solid explanation for the difference in the degree of fitness in reports of different languages. Furthermore, this leads us to another question: whether the translation text can be used as the representative of the target language in corpus research, especially in the study of the verification of the applicability of Zipf's Law. As analyzed above, for text types that contain many social issues and national characteristics, the translated text may not be that suitable to be selected as the representative of the target language. But as for other text types which represent daily life conversation or other common sense, the difference between authentic text and translated text may not be that obvious. Therefore, both the origins of the text and text type should be taken into consideration and viewed as important variables. Thus, further studies can analyze whether there is a significant difference in the applicability of Zipf s Law between translation texts and texts written directly in the target language, with different types of texts being analyzed.

In addition to the above two research questions, this study also found a general phenomenon by observing the curve of fitness degree change. It shows that the middle section of the text significantly fits better Zipf's Law than the beginning and end sections. This finding also corresponds with existing research.

5. Conclusion

This study verified the applicability of Zipf's Law to the report of the 20th National Congress and its multilingual translations, including English, Spanish, and Russian versions. The result of this study gives further evidence to the general applicability of Zipf's Law across different languages and text types. As mentioned above, while there is a large number of research conducted to verify the applicability of Zipf's Law, to our knowledge, the number of studies that focus on report text is still relatively limited. Therefore, the current study filled this gap to a certain extent, specifically in terms of the verification of the applicability of Zipf's Law in the Chinese report text. In addition, the present study can also give some inspiration for further research, as more studies can be conducted to explore the applicability of Zipf s Law in different languages and text types. We can further expand the scope of the research objects by including various texts. For example, texts such as classical Chinese, poems, and scripts. All these types of text have their noteworthy cultural characteristics, and the frequency distribution of words in these texts may be very different from that of spoken texts, so studies on the verification of the applicability of Zipf's Law to these particular texts types are of great value. In addition, this study also has some practical value. In the process of detecting the fitness of the selected texts to Zipf's Law, we have a clearer understanding of the high and low-frequency words in the reports. By analyzing the frequency of words, we can better grasp the latest trends and dynamics of social development. Take the frequency distribution of words in the Chinese report text as an example, the most frequently appeared words in the report include "modernization", "innovation", "characteristics" and so on. These words reveal the current focus of China's development and reflect the main concerns of current national policy as the country firmly pushes ahead with modernization. We can also see in these words the great importance of innovation in the national strategy and the insistence on building a development path with Chinese

characteristics. Based on this, we are inspired that Zipf's Law can also be used as a tool for the detailed analysis and interpretation of report texts.

However, though we have obtained detailed statistical data and basic conclusions from the analysis of these four versions of the reports, there are also some limitations that we must address. Generally, the main limitations in this study are the following three:

Firstly, there exists the question of precision in the word-splitting process conducted on the Chinese text, that is, can the automatic word-splitting program be trusted that it can achieve 100% precision? During our examination of the result of the word-splitting result, we found that some words are split under seemingly ununified criteria, which puts a question mark on the quality of the Chinese text.

Secondly, the size of the corpus may cause a slight deviation in the analysis, as mentioned above in the previous section. Unfortunately, the size of the report cannot be in any way enlarged or reduced to satisfy the need of the study, therefore to avoid the occurrence of similar questions in future studies, the corpus might need to be more carefully chosen.

Thirdly, the underlying mechanisms of the Zipf-Mandelbrot Distribution are still in question. The analysis of the word frequencies of the report text is more or less a descriptive study rather than an analytical one. Because of the loss of the theoretical basis of psychology, the current description of the Zipfian distribution can only be sketched at the mathematical level but not on the mechanistic level, and future studies must dive deeper below the surface of empirical observation of the word frequencies and word frequency ranks.

References

[1] Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley.

[2] Moreno-Sánchez, I., Wilde, M. M., & Corral, Á. (2016). Large-Scale Analysis of Zipf's Law in English Texts. PLOS ONE, 11(1), e0147073.

[3] Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language. Harvard University Press.

[4] Mandelbrot, B. (1953). An informational theory of the statistical structure of language. Communication theory, 486–502.

[5] Wang, Y., & Liu, H. (2022). Revisiting Zipf's law: A new indicator of lexical diversity. In De Gruyter eBooks (pp. 193–202). De Gruyter.

[6] Zipf, G. K. (1935). The psycho-biology of language: An introduction to dynamic philology. Houghton, Mifflin.

[7] Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. Journal of the American Statistical Association, 39(228), 527. https://doi.org/10.2307/2280636

[8] Simon, H. A. (1955). On a class of skew distribution functions. Biometrika, 425–440.

[9] Corominas-Murtra, B., & Solé, R. V. (2010). The universality of Zipf's law. Physical Review E, 82(1). https://doi.org/10.1103/physreve.82.011102

[10] Li, M., & Vitányi, P. M. B. (2019). An Introduction to Kolmogorov Complexity and Its Applications. In Texts in computer science. Springer International Publishing. https://doi.org/10.1007/978-3-030-11298-1

[11] Li, Y., Du, Y., Liu, F., Zhang, Y., Li, M., Wang, J., ... & Yang, Y. (2022). Applicability of Zipf's. The law in Traditional Chinese Medicine Prescriptions. Chinese Medical Sciences Journal, 37(3), 195-201.

[12] Linders, G., & Louwerse, M. M. (2022). Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. Psychonomic Bulletin & Review, 30(1), 77–101. https://doi.org/10.3758/s13423-022-02142-9

[13] Qin, Kexiao. (2020). The Applicability of Zipf's Law in Ancient Chinese Texts—Taking Bibliometrics of Dream Brook Sketchbook as an Example. Shanxi Library Journal, (4), 52-59.

[14] Yu, S., Xu, C., & Liu, H. (2018). Zipf's law in 50 languages: its structural pattern, linguistic. Interpretation, and cognitive motivation. arXiv preprint arXiv:1807.01855.