Higher Order Character Frequency Distribution in Modern Chinese Texts: Application of Zipf's Law

Si Xiaolian

***Abstract*—To investigate the distribution of Chinese characters used in modern Chinese written texts, the higher order character frequency distribution of the *Selected Works of Deng Xiaoping* and *Ordinary World* was researched using Zipf's law. The results show that higher order frequency characters in modern Chinese written texts are consistent with Zipf's law; however, there are a significant number of low-frequency characters. The higher order character frequency distributions are satisfactorily consistent with Zipf's law. Most of the coefficients of determination (R^2) of the fitted straight lines are greater than 0.9, indicating excellent goodness of fit. Character frequency and higher order character frequency distribution patterns have important significance for establishing statistics-based computational language models for modern Chinese.**

***Index Terms*—Zipf's law, character frequency, higher order character frequency, Chinese texts.**

I. Introduction

When studying the real texts of human languages, the Harvard linguist Zipf discovered through extensive statistical analysis that the frequency of a word and its frequency rank (ordinal number) were in an inversely proportional relationship; such a relationship was later named the Zipf's law [1]. Based on the pioneering work of Zipf, researchers in China and other countries successively discussed the word frequency characteristics of various languages; they discovered that many languages were consistent with Zipf's law [2]-[7]. Research on Zipfian distributions has already become a focus in linguistics.

With respect to research on the Chinese language, some researchers discovered that the Chinese character distribution was not strictly consistent with Zipf's law. Wang *et al*.[8] statistically analyzed representative Chinese works from different periods and discovered that there were significant differences in character frequency distribution among written texts from different periods. Guan *et al*.[9] studied the relationships between the frequencies and the frequency ranks of the language units in modern Chinese and found that the relationship between character frequency and frequency rank for each of the 3 different levels of language unit in modern Chinese was the polynomial attenuation function of the Zipf curve. Wang *et al*.[10] studied the character frequency distributions in written texts from different historical periods and attempted to study how humans expressed themselves in Chinese through analyzing the character frequencies; the results showed that the meanings could be expressed in single characters in ancient Chinese, and thus, the character frequency distribution was consistent with Zipf's law; however, there were copious new words in modern Chinese that were composed of simple characters, and thus the character frequency distribution was no longer consistent with Zipf's law.

The aforementioned studies mainly focused on the relationships between the frequencies and the frequency ranks of morphemes. As early as 1992, Sun studied the patterns of the occurrence of same-frequency words [11]. However, there have rarely been any reports on studies on same-frequency words since Sun's study. Inspired by these studies on same-frequency words, in the present study, I further investigated the same-frequency character and higher order character distribution patterns based on the modern Chinese character frequency distribution. In This paper, the first order and second order character frequency represent the traditional character frequency and same-frequency character frequency, the third order character frequency means the same-same-frequency character frequency, as shown in below.

II. Data and Methods

To compare the character frequency distribution in works by the same author from different times or in different volumes of the same work, we obtained two literary works, *Selected Works of Deng Xiaoping* (*Deng*, hereinafter) and *Ordinary World* (*Ordinary*, hereinafter), from the Internet gratis and reorganized the two works. In addition, we divided each of the two literary works into 3 volumes (volumes 1, 2 and 3), and used the 3 volumes of each work together with the complete set (comprising all 3 corresponding volumes) of each of the two literary works as the study objects. *Deng* is selected literary works and speeches by Comrade Deng Xiaoping from different times; volumes 1, 2 and 3 of *Deng* correspond to the time periods of the 28 years before the Cultural Revolution, 1975-1982 and 1982-1992, respectively. The 3 volumes of *Deng* span a large time frame, and there is no overlap among the 3 volumes in terms of time. *Ordinary* is a great, full-length novel with a total of over a million words by Lu Yao. Lu comprehensively revised the novel after finishing the first draft. Thus, the volumes of *Ordinary* were generally completed during the same time period.

Prior to data analysis, any non-Chinese symbols such as the punctuation marks, English letters and Arabic numerals in the selected works were completely removed. In addition, the chapter titles were also removed. The computer-aided statistical analysis of the character frequencies was completed using an application programmed in MATLAB.

Manuscript received September 23, 2018; revised November 22, 2018.

Si Xiaolian is with the College of Chinese language and Literature, Northwest Normal University, Lanzhou, Gansu 730070, PR China (e-mail: sixiaolian1979@163.com).

doi: 10.18178/ijlll.2018.4.4.186

According to fractal theory, the ranking number of a character's frequency represents the total number of different characters that is equal to or greater than that character's frequency. Therefore, after they had been numbered, the Chinese character frequencies were sorted from the highest to lowest; same character frequencies were numbered with only one ranking number, which was the ranking number of the last Chinese character of the same-frequency characters.

According to Zipf's law, the relationship between character frequency and rank in a relatively long article (with a total of at least approximately 5,000 words) satisfies the following rule:

$$p=C(r)^{-\beta} \quad (1)$$

where p represents the character frequency that is ranked in the r*th* position, β represents the Zipf fractal dimension, the similar dimension D=1/β, and C is a constant. We take the logarithm of both sides of the above equation:

$$\lg(p)=C'-\beta\lg(r) \quad (2)$$

where C'=lg(C). lg(p) is plotted along the ordinate (y-axis), and lg(r) is plotted along the abscissa (x-axis). If the character frequency is consistent with Zipf's law, then the curve should be a straight line. The curve is then fitted with a straight line. The intercept of the straight line on the y-axis is C', and the slope of the straight line is -β.

III. Results and Discussion

A. First Order Character Frequency Distributions

Table I lists the total numbers of Chinese characters, number of different Chinese characters and mean number of occurrences of different Chinese characters in *Deng* and *Ordinary*. The total number of characters in volumes 1 and 3 of *Deng* are almost same, which are slightly lower than the number of characters in volume 2; however, the numbers of different Chinese characters in all 3 volumes are basically the same (there are approximately 2,000 different Chinese characters in each volume), indicating that the number of different Chinese characters that Deng Xiaoping used basically remained the same in different time periods. The number of different Chinese characters in the complete set of *Deng* is approximately 1.25 times the number of different Chinese characters in each volume, indicating that Deng used different Chinese characters in different time periods, i.e., some characters were only used in certain time periods; as time went by, some characters were forgotten, whereas other characters were used again. The total numbers of characters in each of the 3 volumes of *Ordinary* are basically the same. However, there are relatively large differences in the number of different Chinese characters among the 3 volumes of *Ordinary*. In particular, the number of different Chinese characters is the highest in volume 2, and volume 2 includes all of the Chinese characters that are used in the complete set of *Ordinary*, which is attributed to the fact that Lu Yao completed *Ordinary* within a very short time period. Due to the author's writing habits and memories, the characters that he had used before would be easily used again; as the writing continued, the characters that Lu Yao mastered had been basically all used, and in the later stages of writing, Lu basically repeatedly used the characters that he had used earlier.

TABLE I: Statistical Data of Deng and Ordinary

Works	Total numbers of Chinese characters	Number of different Chinese characters	Mean number of occurrences of different Chinese characters
Complete set of *Deng*	536318	2574	208.4
Volumes 1 of *Deng*	169043	2030	83.3
Volumes 2 of *Deng*	207632	2039	101.8
Volumes 3 of *Deng*	159643	1984	80.5
Complete set of Ordinary	704945	3728	189.1
Volumes 1 of *Ordinary*	238286	3066	77.7
Volumes 2 of *Ordinary*	231791	3728	62.2
Volumes 3 of *Ordinary*	234868	3269	71.8

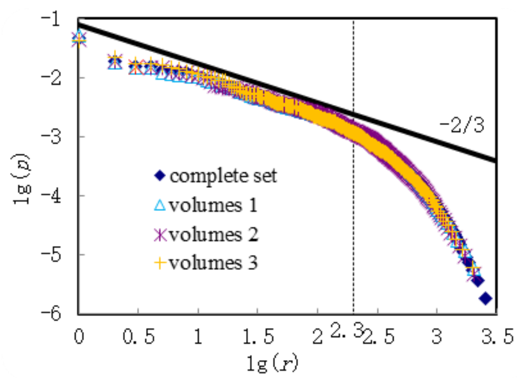

Fig. 1. Character frequency in *Deng*.`2QYT------------

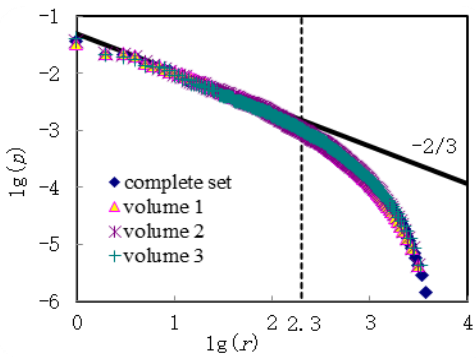

Fig. 2. Character frequency in *Ordinary*.

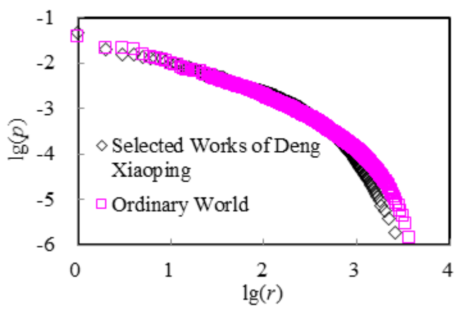

Fig. 3. Character frequency in complete sets of *Deng* and *Ordinary*.

Fig. 1, Fig. 2 and Fig. 3 show the character frequency distributions in *Deng* and *Ordinary*. It can be observed in Fig. 1 and Fig. 2 that the character frequency distribution curves of the complete sets of *Deng* and *Ordinary* and their volumes are very similar (all of which are convex curves), which is consistent with You's findings [12] and indicates that there is memory, to some extent, in the usage of Chinese characters in different works by the same author. Specifically, there are significant deviations for the 5 Chinese characters with the highest character frequencies in the two works from the overall curves. The distribution curves between ranking numbers 6 and 200 ($10^{2.3}$) are almost straight lines. Each of the 4 straight lines of the two works is almost parallel to the other 3, and the straight lines all have a slope of -2/3. When the ranking number is greater than 200, the curves start descending; the character frequencies decrease rapidly with the increasing ranking number, indicating that there are small numbers of low-frequency words, and the curves are no longer consistent with the straight line pattern, which is inconsistent with the word-frequency distribution patterns of Chinese and other languages. However, this phenomenon also further proves that the basic language unit of Chinese is the word. Chinese words consist of monosyllabic words and disyllabic words, and most Chinese words are disyllabic words; therefore, there are a relatively small number of low-frequency words. It can be observed in Figure 3 that there are relatively large differences in the number of characters among the low-frequency characters, particularly the characters with a frequency of less than 200 ($10^{2.3}$) within the complete sets of the two works, whereas the distribution patterns of the characters with a frequency of over 200 are almost the same, which is because *Deng* includes the speeches that Comrade Deng Xiaoping gave during different time periods, and most of the speeches consist of spoken Chinese words; thus, the number of low-frequency characters is lower than in written texts.

B. Second Order Character Frequency Distributions

The numbers of same-frequency characters in the literary works (p_2) were statistically analyzed using the MATLAB software. The p_2 were then sorted from highest to lowest according to its values. Using the same notation method used for the character frequencies, for the same values of p_2, only the last position was noted (r_2). Similarly, $\lg(p_2)$ is plotted along the ordinate (y-axis), and $\lg(r_2)$ is plotted along the abscissa (x-axis). Thus, the same-frequency character distribution curves were obtained to study their distribution patterns. We call same-frequency character distributions as second order character frequency distributions. Figures 4 and 5 show the second order character frequency distributions in the two works. It can be observed in Figures 4 and 5 that there are good linear relationships between the logarithms of the numbers of same-frequency characters in the two works and the logarithms of the corresponding ranking numbers; Table II lists the slopes and intercepts of the fitted straight lines. It can be observed in the table that the coefficients of determination (R^2) are all greater than 0.988, indicating that the goodness of fit of each straight line that fits the actual data points is excellent, and all of the secondary character frequency satisfy a Zipfian distribution.

Specifically, the β values of the complete sets of two works and their volumes are very close to 1, indicating that the secondary character frequency distributions strictly obey the relationship $p_2r_2=C$. However, further studies are necessary to understand the internal cause of this phenomenon. It can be observed in the C values listed in Table II that the C values of the complete set of *Deng* and its volumes are basically the same (approximately 350); similarly, the C values of the volumes of *Ordinary* are also basically the same as each other (approximately 700). However, they are slightly greater than the C value of the complete set of *Ordinary*, indicating that the same-frequency distribution patterns in the different parts of the same work are also basically the same.

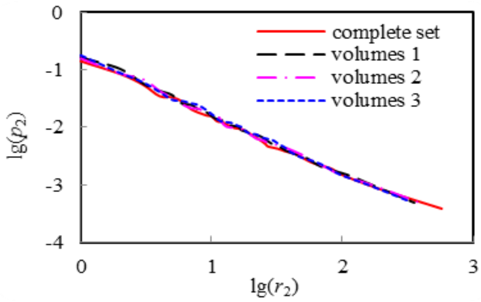

Fig. 4. Secondary character frequency distributions in *Deng*.

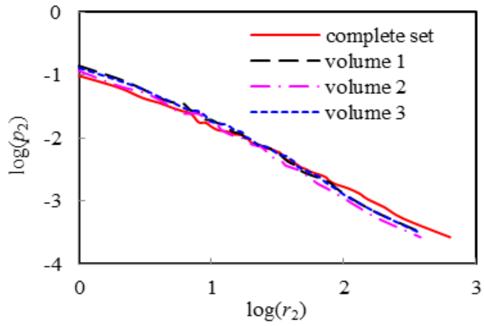

Fig. 5. Secondary character frequency distributions in *Ordinary*.

TABLE II: PARAMETERS OF ZIPF'S LAW OF SECOND ORDER CHARACTER FREQUENCY IN *DENG* AND *ORDINARY*

Works	C'	$C=10^{C'}$	β	R^2
Complete set of *Deng*	-0.8457	0.1427	0.9818	0.9946
Volumes 1 of *Deng*	-0.7629	0.1726	1.0247	0.9971
Volumes 2 of *Deng*	-0.7928	0.1611	1.0063	0.9951
Volumes 3 of *Deng*	-0.7483	0.1785	1.0282	0.9977
Complete set of Ordinary	-0.8478	0.1420	0.9522	0.9929
Volumes 1 of *Ordinary*	-0.6606	0.2185	1.0791	0.9915
Volumes 2 of *Ordinary*	-0.7260	0.1879	1.0829	0.9887
Volumes 3 of *Ordinary*	-0.6594	0.2191	1.0746	0.9884

C. Higher Order (Third Order) Character Frequency Distributions

We repeated the statistical analysis that had been conducted in Section 2.2 to continuously statistically analyze the frequencies (p_3) of the occurrence of p_2 in the secondary character frequency distributions. The p_3 were then sorted based on its values and numbered with ranking numbers (r_3). And thus, new relationships between the frequencies and ranking numbers were obtained. The third order character frequency distribution curves of the works could be obtained by plotting $\lg(p_3)$-$\lg(r_3)$ diagrams on coordinate paper (Fig. 6 and Fig. 7). It can be observed in Figures 6 and 7 that the third order character frequency distributions in the works are consistent with Zipf's law.

Table III list the related parameters of the fitted straight lines of the curves in Fig. 6 and Fig. 7. It can be observed in Fig. 6 and Fig. 7 and Table III that the coefficients of determination (R^2) of the fitted straight lines of the third order character distributions in the two works are all greater than 0.96, indicating excellent goodness of fit. The β and C values of each volume of each work are close to the β and C values of the other two volumes, respectively. However, there are relatively large differences between the parameters for the complete set of each work and its volumes, which may be because the C value is related to the total number of characters based on the preliminary analysis.

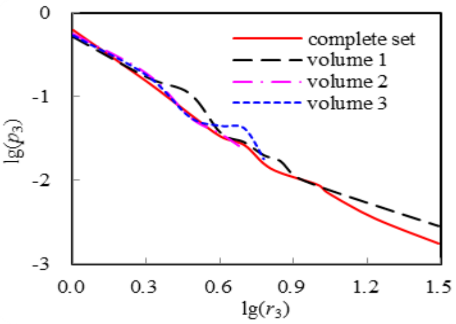

Fig. 6. Tertiary character frequency distributions in *Deng*.

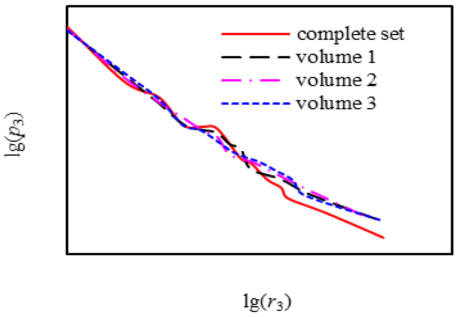

Fig. 7. Tertiary character frequency distributions in *Ordinary*.

TABLE III: PARAMETERS OF ZIPF'S LAW OF THIRD ORDER CHARACTER FREQUENCY IN *DENG* AND *ORDINARY*

Works	C'	$C=10^{C'}$	β	R^2
Complete set of *Deng*	-0.3555	0.4410	1.7159	0.9844
Volumes 1 of *Deng*	-0.3590	0.4375	1.6126	0.9702
Volumes 2 of *Deng*	-0.3771	0.4197	1.6066	0.9743
Volumes 3 of *Deng*	-0.3836	0.4134	1.5558	0.9639
Complete set of Ordinary	-0.3790	0.4179	1.5829	0.9813
Volumes 1 of *Ordinary*	-0.4442	0.3596	1.4440	0.9760
Volumes 2 of *Ordinary*	-0.4431	0.3605	1.4447	0.9825
Volumes 3 of *Ordinary*	-0.4451	0.3588	1.4349	0.9780

IV. CONCLUSIONS

In the present study, the character frequency distribution patterns in the complete sets of two literary works as well as within 3 volumes subdividing each literary work were used as the experimental objects of study. In contrast to the previous commonly used methods for discussing character frequency distributions, we hoped to investigate deeper character frequency distribution patterns through studying second and third order character distribution patterns. The character frequency distributions of high-frequency characters, particularly characters with a frequency of over 200 ($10^{2.3}$) are basically consistent with Zipf's law. However, the character frequency distributions of the 5 Chinese characters with the highest character frequencies are significantly inconsistent with Zipf's law, which is closely related to the usage patterns of these Chinese characters. The logarithm of the character frequency of each of the Chinese characters with a frequency of less than 200 is no longer in a linear relationship with the logarithm of the ranking number, indicating that there are a small number of low-frequency characters. The character distribution curve of the complete set of each work is almost parallel to the character distribution curves of the 3 volumes of the work, indicating that the numbers of different Chinese characters that an author used in different time periods are basically the same and that the usage patterns of characters in different volumes of the same work from the same author are also basically the same.

The statistical analysis of the experiment shows that the second and third order character distribution patterns in written texts are consistent with Zipf's law. The coefficients of determination (R^2) of the fitted straight lines are basically all greater than 0.9, and thus the goodness of fit is excellent, indicating that multi-level character distributions in modern Chinese written texts exhibit significant fractal characteristics. Thus, we believe that the character frequency distribution patterns in modern Chinese texts are relatively deeply related to the way in which humans express themselves in Chinese. Therefore, we should pay attention to higher order character frequency distribution patterns when establishing statistics-based computational language models for modern Chinese.

REFERENCES

[1] G. K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, London: George Routledge & Sons Ltd., 1936.

[2] N. Hatzigeorgiu, G. Mikros, and G. Carayannis, "Word length, word frequencies and Zapf's law in the Creek language," *Journal of Quantitative Linguistics*, vol. 3, pp. 175-185, 2001.

[3] S. W. Choi, "Some statistical properties and Zipf's law in Korean text corpus," *Journal of Quantitative Linguistics*, vol. 1, pp. 19-30, 2000.

[4] R. D. Smith, "Investigation of the Zipf-plot of the extinct meroilic language," *Glottometrics*, vol. 15, p. 53, 2007.

[5] B. D. Jayaram and M. N. Vidya, "Zipf's law for Indian languages," *Journal of Quantitative Linguistics*, vol. 4, pp. 293-317, 2008.

[6] L. Q. Ha, D. Stewart, P. Hanna, and F. Smith "Zipf and type-token rules for the English, Spanish, Irish and Latin languages," *Web Journal of Formal Computational & Cognitive Linguistics*, vol. 8, pp. 1-12, 2007.

[7] W. Li, "Fitting Chinese syllable-to-character mapping spectrum by the beta rank function," *Physica A: Statistical Mechanics and its Applications*, vol. 4, pp. 1515-1518, 2012.

[8] D. Wang, M. Li, and Z. Di, "True reason for Zipf's law in language," *Physica A.*, vol. 358, pp. 545-550, 2005.

[9] Y. Guan, X. Wang, and K. Zhang "The frequency-rank relation of language units in modern Chinese computational language model," *Journal of Chinese Information Processing*, vol. 2, pp. 8-15, 1999.

[10] Y. Wang, Y. Liu, and Q. Ceng "Zipf distribution of word use in Chinese literature," *Journal of Beijing Normal University*, vol. 4, pp. 424-427, 2009.

[11] Q. Sun, "Calculation method of number of words having the same frequency occurring in the text-The development of Zipf's second law," *Journal of Northeast Normal University*, vol. 2, pp. 34-39, 1992.

[12] R. You, "Zipf's law and the distribution of Chinese character frequency," *Journal of Chinese Information Processing*, vol. 3, pp. 60-65, 2000.

Si Xiaolian was born in Lintao county, Gansu province, China in 1979. She received her bachelor of arts from Northwest Normal University and has received both her master of arts and doctor of literature from Beijing Normal University in 2005 and 2009, respectively.

She is presently an associate professor at College of Chinese language and Literature, Northwest Normal University. Her research interests include philology, linguistic and language.