International Journal of Knowledge and Language Processing
Volume 12, Number 1, **2021**

www.ijklp.org
KLP International ©2021 ISSN 2191-2734
pp.22–34

A Study of the Differences between the Function Words and Content Words in Modern Chinese Based on the Parameters of Word Frequency Distribution

Ying Zhang[12] and Pengyuan Liu[12]

[1]School of Information Science, Beijing Language and Culture University
[2]National language resources monitoring and Research Center for print media
No.15, Xueyuan Road, Haidian District, Beijing, China
zhangyingblcu@163.com, Corresponding author: liupengyuan@pku.edu.cn

Received August 2020; revised September 2020

ABSTRACT. *According to the syntactic function of words in sentences and whether they express real meanings, words can be divided into two categories: content words and function words. The main function of content words is to express real meanings, while function words mainly play the syntactic function in sentences. However, there is still no accurate distinction between function words and content words, and many scholars have different opinions on the attribution of the part of speech in modern Chinese. Therefore, this paper attempts to use the parameters of word frequency distribution in quantitative linguistics, that is, Zipf's law to fit the function words and content words extracted from 18 texts of different sizes and styles. It is found that both function words and content words conform to Zipf's distribution, and their parameter values are significantly different. On the distribution curve, they also show obvious differences: the content words tend to be flat, while the function words are concentrated in the first half of the word frequency distribution, which shows a rapid downward trend.*

Keywords: Zipf's Law; Parts of Speech; Function Words; Content Words

1. **Introduction.** The division of modern Chinese parts of speech into content words and function words has always been a problem for linguists. Scholars hold different opinions on whether some parts of speech should be divided into function words or content words. The division of function words into function words and content words is the basis of modern Chinese studies, which is of certain help to the study of more applied linguistic knowledge.

The law of word frequency distribution in quantitative linguistics——Zipf's law describes the quantitative characteristics of language structure in language systems and language use, and is applicable to the study of different language units. At present, no scholar has explored and verified the distinction between parts of speech in modern Chinese by the method of Quantitative Linguistics. Therefore, this article explores the relationship between words and word frequencies in both type of words through a comparative study of different scales and different styles. We try to use data to analyze the difference between function words and content words in modern Chinese parts of speech, and lay the foundation for the subsequent research on the issue of the controversial part of speech attribution.

2. **Related Work**.

2.1. **The Division of Parts of Speech**.

The study of Chinese parts of speech has a long history. As early as the Han Dynasty in China, there was the study of "auxiliary word". *ZhuYuCi*, written by Lu Yiwei in Yuan Dynasty, is the first time to focus on Chinese function words. The concept of parts of speech was put forward in the Qing Dynasty, whose representative works are Liu Qi's *Zhu Zi Bian Lue* and Wang Yinzhi's *Jing Zhuan Shi CI*.

Ma's *Ma Shi Wen Tong*[1] is the first systematic work of Chinese grammar research.In terms of part of speech, compared with the Latin grammatical system, it is divided into two categories: real characters and virtual characters, and then specifically divided into nine categories, establishing the part of speech system of Classical Chinese for the first time. Li's *New Chinese Grammar*[2] follows Nasfield's *English Grammar* and divides Chinese words into five categories and nine types, establishing a complete Mandarin vernacular parts-of-speech system.

Many scholars have also put forward their own views on the division of parts of speech in their continuous research. Ding[3] directly divides parts of speech into eleven categories, without dividing the function words and the content words. Lu[4] and Zhu[5] divide Chinese words into two parts: the "closed classes" and the "open classes". They have a deeper understanding of parts of speech and propose that parts of speech are linked with other grammatical problems. Wang[6] puts forward the theory of semi-notional words and semi-functional words. Because some words in Chinese are between content words and function words in terms of meaning and function, it is difficult to directly determine which category they belong to.

With the broadening of scholars' research horizon and the continuous exploration of research methods, Yuan[7], from the perspective of cognition, believed that the division should be based on the similarity of word categories, first dividing words into "single word" and "double word", and then dividing words into content words and function words from "single word". Based on computational linguistics, Guo[8] proposes the classification of the expression function of words at the lexical level, and divides Chinese vocabulary into 19 categories, which separate compound words and independent words, and separate content words and function words in compound words.

2.2. **Research on Parts of Speech Using Quantitative Linguistics.** Zipf's law, one of the earliest statistical laws in quantitative linguistics, is about the frequency distribution of words in texts. It is found that the product of word frequency and frequency order generally stable at a constant K. Zipf's law have strong applicability in different languages and different language units.

In the research on the applicability of Zipf's law to a single language, a wide range of languages are involved, such as Chinese (Wang et al.[9] ; He[10]), English (Williams et al.[11]), Spanish, Irish and Latin (Ha[12]), Greek (Hatzigeorgiu[13]), Hindi (Jayaram[14]), Korean (Choi[15]), Turkish (Dalkilic & Cebi[16]), Italian (Tuzzi et al.[17]), etc. Among the studies on different linguistic units by Zipf's law, there are phonemes (Xin et al.[18]), syllables and word frequency (Ha et al.[19]), word frequency (Chen et al.[20]; Wang Yang et al.[9]), and some units larger than word frequency, such as N-gram structures (Guan et al.[21]; Ha et al.[19]), phrases (Williams et al.[11]), etc. Most of these studies are about the language unit of word.

Regarding the study of parts of speech, Piantadosi[22] used the Penn Treebank (Marcus et al.,[23]) syntactic categories (such as qualifiers, nouns, third-person singular verbs, etc.) to analyze the frequency distribution and homogeneity of each syntactic category in the Brown Corpus. The word frequency distribution of words in the syntactic category found that the frequency distribution of the syntactic category conforms to Zipf's law, but the applicability of the word frequency distribution under different syntactic categories to Zipf's law is different, and the Zipf parameters of the specific fitting are different. The whole is in accordance with Zipf distribution.

From the perspective of application value, Liu[24] said language research related to Zipf law is not only of great significance to computational linguistics, language information processing, corpus linguistics, language teaching and testing, but also some parameters in Zipf's law to be used as indicators of language classification.

Under the guidance of different theories, there are different standards for the division of content words and function words. Therefore, this paper considers whether we can find out the characteristics of function words and content words in the frequency distribution parameters based on the current classification, so as to quantify the differences between them.

3. Data Preparation.

3.1. **Data Selection**. We select five types of language styles: news, subtitles, microblog, literary works and online novels. News corpus is from the Dynamic Circulation Corpus (DCC). Because the size of the text in the corpus is large enough, four texts of different sizes are randomly selected from the corpus, covering million words of text. The microblog corpus is from Sina Weibo's microblog in January 2013, and the subtitle corpus is from the Chinese translation subtitles of many American dramas. Three texts of different sizes are randomly selected from them. Literary works are selected from the works of three traditional literary writers, respectively covering thousands of words, ten thousand words and one hundred thousand words. There are many categories of online novels, and five

texts with 1.5 million words are randomly selected. Therefore, the selected text covers five different styles and reaches the scale from 1000 words to 100 million words. The details are shown in Table 1:

TABLE 1. THE SCALE OF THE SELECTED CORPUS.

Styles	Type	token	Styles	Type	token
DCC1	169071	6840873	subtitle1	125287	6101401
DCC2	304059	17544710	subtitle2	203288	20730004
DCC3	575191	87943002	subtitle3	261304	37828169
DCC4	781825	162359011	microblog1	212231	5596161
online novels1 (gay fictions)	185719	13776858	microblog2	355065	15679279
online novels2 (urban life)	220486	17456720	microblog3	856315	71455964
online novels3 (entertainers)	168935	17462041	Literary works1	1265	3373
online novels4 (school youth)	210286	17609620	Literary works2	9214	44422
online novels5 (supernatural fiction)	255837	17629048	Literary works3	16692	119589

3.2. **Data preprocessing.** Use the jieba word segmentation tool to perform word segmentation and part-of-speech tagging on the selected text and the marked parts of speech are extracted according to the categories in Table 2:

TABLE 2. PATRS OF SPEECH CATEGORY.

Type	Category
Content words	Noun/n Verb/v Adjective/a Numeral/m Quantifier/q Distinguishing word/b Pronoun/r Place word/s Time word/t Location word/f
Function words	Preposition/p Auxiliary word/u Conjunction/c

3.3. **Data processing.** We fitted Zipf's law $y = ax^{(-b)}$ to each vocabulary of each text, and then took the logarithm of the frequency order and frequency and added 1 (he x-axis is [log(frequency sequence)+1], the y-axis is [log(frequency)+1]. The discrete data points approximate a straight line, so the linear formula $y=-ax+b$ was used for the data fitting.

Take the noun (N) vocabulary extracted from the DCC news text as an example, each sub category word list is arranged from large to small according to frequency, and frequency order is added, as shown in Table 3:

TABLE 3. THE SITUATION OF THE SELECTED CORPUS.

Frequency sequence	Word	Frequency	Frequency sequence	Word	Frequency
1	中国	608801	6	社会	34367
2	人	57028	7	建设	31995
3	发展	55641	8	国家	30059
4	工作	48284	9	企业	28179
5	问题	37604	10	经济	25811

Then we used formula $y = ax^{(-b)}$ to fit the frequency sequence and frequency data of the vocabulary, as shown in Figure1. The fitting result of Zipf's law is: $y = 25067174.35x^{(-1.444)}$, the goodness of fit $R^2 = 0.977$, which shows that the fitting result is very good and conforms to the Zipf distribution.

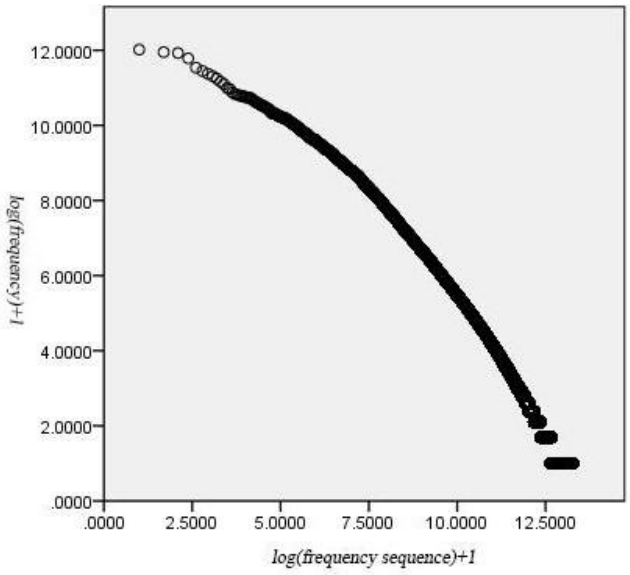

FIGURE 1.ZIPF'S LAW FITTING OF NOUN WORD LIST IN DCC NEWS TEXT.

Then a linear fitting was performed. As shown in Figure2, the linear fitting result is: $y = -1.444x + 9.989$, the goodness of fit $R^2 = 0.977$, indicating that the fitting result is very good .

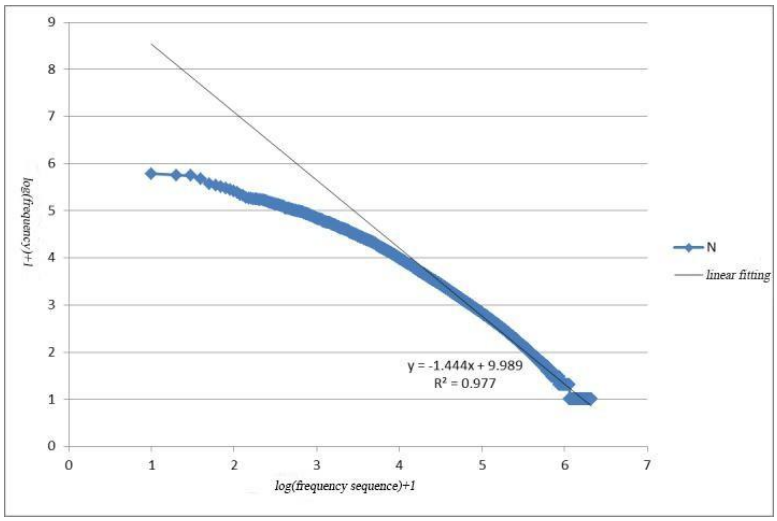

FIGURE 2. LINEAR FITTING OF NOUN WORD LIST IN DCC NEWS TEXT.

4. Data Analysis.

4.1. **The fitting of Zipf's law**. Through the statistical analysis of each word list from all

corpus texts (Table 4), it is found that the average value is 0.88857, the minimum value is 0.603, and the overall fitting result is good. It can be said that Zipf's law better describes the word frequency distribution data of different vocabularies of the text. Zipf's Law, which is used to describe the distribution of word frequency, is universally verified in Chinese texts of different parts of speech.

TABLE 4. DESCRIPTIVE STATISTICS OF POWER LAW GOODNESS OF FIT

	Sample	Minimum	Maximum	Mean value	Standard deviation
R^2	162	0.603	0.984	0.88857	0.097008

While the word frequency distribution data of the vocabularies of these different texts follow the same distribution law, are there any significant differences? The following is a statistical analysis of the parameter b of Zipf distribution in the fitting results. The average value of Zipf distribution parameter b for different parts of speech is shown in Table 5: the total vocabulary is 1.48106 and the content words (1.469) is lower than the function words (2.58972). However, whether the difference of parameter b among different sample groups is statistically significant needs further testing.

TABLE 5. DESCRIPTIVE STATISTICS OF POWER LAW INDEX FOR DIFFERENT TEXT VOCABULARIES

	Sample size	Mean value	Standard deviation
All words	18	1.49106	0.310474
Content words	18	1.46900	0.330000
Noun	18	1.37672	0.313630
Verb	18	1.64544	0.394908
Adjective	18	1.72744	0.472299
Function words	18	2.58972	0.400968
Preposition	18	2.29222	0.280182
Conjunction	18	2.40550	0.533929
Auxiliary word	18	3.94311	0.518898

The independent sample T-test is conducted with the zipf distribution parameter b as the independent variable and the function words and the content words as the independent variables. The homogeneity test results of variance are as follows: $F(1,36)=0.89$, $P=0.768$. The Zipf parameter b fitted by function words and content words in different texts has a significant difference,$t(36)=9.156$,$p=0.000<0.05$, and the parameter b of function words is greater than that of content words. As shown in Figure 3 below, the size of parameter b of function words and content words is arranged from small to large according to the text size. On the whole, the value of parameter b of function words and content words increases with the increase of text size, and function words are higher than content words.

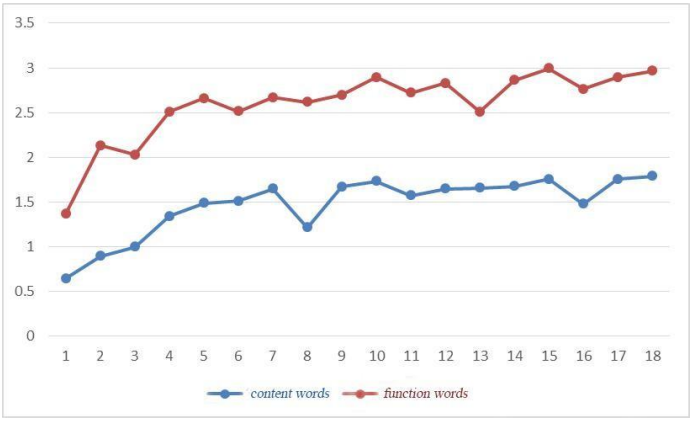

FIGURE 3.THE CHANGE TREND OF THE POWER LAW INDEX OF THE FUNCTION WORDS AND CONTENT WORDS OF EACH TEXT.

In order to further understand whether there are significant differences between function words and content words among texts of different styles and sizes, we first compare whether there are differences between them extracted from each type of text, and then control the factors of style and text size respectively, and conduct one-way ANOVA. The following results are analyzed at 95% confidence level. The contrast of content words consists of three typical parts of speech: noun, verb and adjective; the contrast of function words is composed of three typical parts of speech: auxiliary word, preposition and conjunction.

In the 18 texts, the T-test is performed on the function words and content words of each text. The result is shown in Table6. There are significant differences between the function words and the content words of each text.

TABLE 6.STATISTICAL RESULTS OF THE DIFFERENCE BETWEEN FUNCTION WORDS AND CONTENT WORDS IN 18 TEXTS.

text	Levene's Test		T-test			text	Levene's Test		T-test		
	F	Sig.	t	df	Sig.(two-tailed)		F	Sig.	t	df	Sig.(two-tailed)
1	5.398	0.059	-2.805	6	0.031	10	4.778	0.071	-3.768	6	0.009
2	3.817	0.099	-3.424	6	0.014	11	4.491	0.078	-2.937	6	0.026
3	4.586	0.076	-4.842	6	0.003	12	4.123	0.089	-4.034	6	0.007
4	4.704	0.073	-3.021	6	0.023	13	5.229	0.062	-2.587	6	0.041
5	2.906	0.139	-4.413	6	0.005	14	3.561	0.108	-3.126	6	0.020
6	3.960	0.094	-2.643	6	0.038	15	2.462	0.168	-3.197	6	0.019
7	4.979	0.067	-3.187	6	0.019	16	3.878	0.096	-3.002	6	0.024
8	3.472	0.112	-3.087	6	0.021	17	3.215	0.123	-2.620	6	0.040
9	4.576	0.076	-2.987	6	0.024	18	3.541	0.109	-2.837	6	0.030

In terms of style, a text with a scale of about 17 million words is selected from each of the four types of news, microblog, subtitles and online novels. Due to the small scale of literary works, it does not participate in the comparison. The result of one-way ANOVA of the content words is F=1.103, P=0.386>0.05, that is, there is no significant difference between the parameter b of the content words extracted from the four styles. The result of one-way ANOVA of function words is F=0.092,P=0.963>0.05, that is, there is no significant difference between the parameter b of function words fitting extracted from these four styles. In general, the style has no significant influence on the parameter b between the content words and function words in each text, but there are also differences in the use of parts of speech in each text. As shown in figure 4, we compare the parameter b fitted by the total vocabulary of each text, subtitle > news > online novels > microblog. In content words, the smallest parameter b is microblog; in function words, the smallest parameter b is online novels.

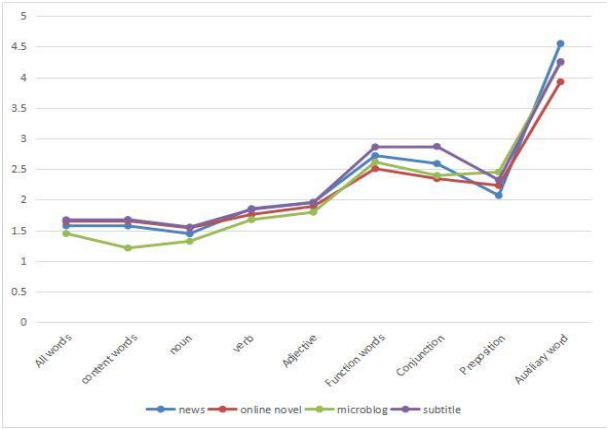

FIGURE 4. POWER-LAW PARAMETER DISTRIBUTION OF FUNCTION WORDS AND CONTENT WORDS AND THEIR TYPICAL PARTS OF SPEECH IN DIFFERENT STYLES OF TEXT.

In terms of text scale, the text scale of literary works is respectively thousands, tens of millions and hundreds of thousands of words, while other texts are in the level of millions, tens of millions or hundreds of millions of words. In literary works, there are significant differences in the parameter b (F=40.541, P=0.000<0.05) of content words of different scales. With the expansion of the text size, the value of parameter b also increases. However, there is no significant difference in the parameter b (F = 1.099, P = 0.374 > 0.05) of function words, but Theatrical Performances < Fortress Besieged < Red Sorghum. In microblog texts, the parameter b of content words (F = 0.698, P = 0.532>0.05) and the parameter b of function words (F = 0.048, P = 0.953>0.05) of texts of different sizes have no significant differences. In the subtitle text, there is no significant difference between the parameter b of text content words of different sizes (F=2.875,P=0.108>0.05) and the parameter b of function words (F=0.234, P=0.796>0.05). In news texts, there is no significant difference between the parameter b of text content words of different sizes (F=1.914, P=0.181>0.05) and the parameter b of function words (F=0.205, P=0.891>0.05).

Van [25] mentioned that the measurement of Zipf rate depends on the size of the corpus. Bernhardsson [26] pointed out that the exponent described by the power law of word frequency distribution seems to change with the length of the text, rather than being constant. From the perspective of this article, the parameter b of function words and content words increases with the expansion of the text size, but in texts with less than 100,000 words, content words have significant differences, which may also be related to the writing style and era of the three text authors. More detailed research can be conducted on small-scale texts.

4.2. **Linear fitting.** Since the power law curve cannot observe the word frequency from the image, the difference between the two distributions can be observed intuitively through linear fitting. According to the statistical analysis of goodness of fit (Table 7), the average value is 0.88856, the minimum value is 0.603, and the fitting result is good on the whole. It can be said that the regression effect of Zipf logarithmic curve distribution is good, which describes the word frequency distribution data of different word lists in the text.

TABLE 7. DESCRIPTIVE STATISTICS OF GOODNESS OF LINEAR FITTING

	Sample size	Minimum	Maximum	Mean value	Standard deviation
R^2	162	0.603	0.987	0.88856	0.097173

The following is a statistical analysis of slope a in the linear fitting results. The average value of the slope of different parts of speech linear fitting is shown in Table 8. The mean value of the linear fitting slope a of the total word table is 1.48106, and the content words(1.469) is lower than that of the function words (2.58972). However, whether the difference in slope among different groups is statistically significant needs to be further tested.

TABLE 8. DESCRIPTIVE STATISTICS OF THE SLOPE OF LINEAR FITTING OF DIFFERENT TEXT VOCABULARIES

	Sample size	Mean value	Standard deviation
All words	18	1.49106	0.310474
Content words	18	1.46906	0.330036
Noun	18	1.37678	0.313663
Verb	18	1.64544	0.394908
Adjective	18	1.72744	0.472299
Function words	18	2.59006	0.401107
Preposition	18	2.29200	0.280148
Conjunction	18	2.40561	0.540029
Auxiliary word	18	3.94217	0.518703

Taking the slope a of the linear fitting as the independent variable, and taking the function word and the content word as the independent variable, the independent sample T-test has been performed. The result of the homogeneity of variance test is: $F(1,36)=0.66$, $P=0.422$; at the 95% confidence level, part of speech factors have a significant impact ($t(36)=6.22$, $P=0.000<0.05$), that is, there are significant differences in the slope a of the linear fitting between the function word and the content word in different texts. The slope a of the function words is greater than that of the content word. The size of the slope a of the function word and the content word generally increases with the size of the text, and the function word is higher than the content word.

By observing the following linear fitting figure 6-9, the distribution curve of the content words is in a steady downward trend. Except for the slight fluctuations of the first few high-frequency words, the rest are steadily declining, with relatively small fluctuations. The first half of the function word curve shows a downward trend, the second half shows a upward convex trend, and the whole shows a rapid downward trend. Comparing the two different fitting methods, it can be seen that content words and function words can fit Zipf's law, but there are differences in parameters, this is similar to Piantados's finding [22] that the frequency distribution of syntactic categories in English (such as determiners, nouns, third-person singular verbs, etc.) conforms to Zipf's law. The differences between different parts of speech in Chinese can also be compared through parameters.

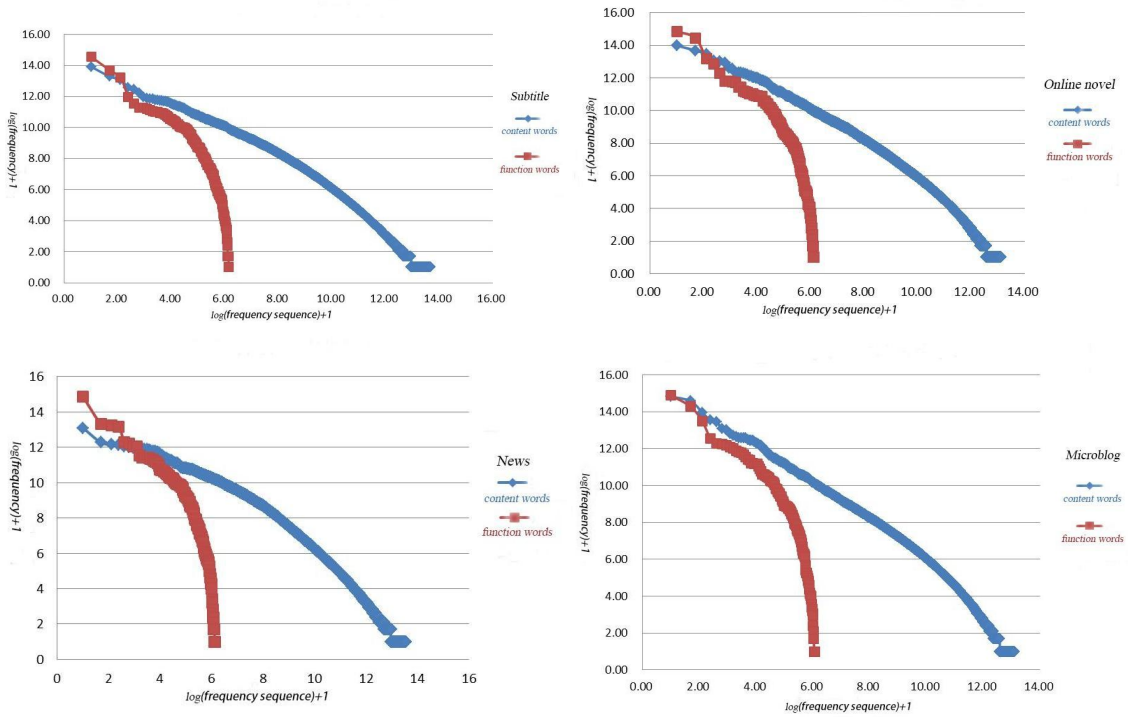

FIGURE 5-8. LINEAR FITTING CONTRAST BETWEEN FUNCTION AND CONTENT WORDS

5. **Discussion.** This article extracts total 162 vocabularies, content words (and three typical part-of-speech nouns, verbs, and adjectives), and function words (and three typical

part-of-speech prepositions, conjunctions, and auxiliary words) from 18 texts of different styles and sizes. Firstly, the Zipf's law is fitted, and the universality of word frequency distribution in different parts of speech Chinese texts is verified. Moreover, there are significant differences in the parameter b of function words and content words fitting between texts and within texts. The average value of the parameter b of the function words is 2.5, and the content words is 1.4. In addition, it is also divided into different styles to conduct analysis, and it is found that there is no significant difference in parameter b of the word fitting between different styles. Among different text sizes, text with content words below 100,000 words has significant difference, while text with content words above 100,000 words has no significant difference. Function words are not affected by text size.

Secondly, the frequency sequence and frequency are taken to logarithm and then linearly fitted. Han et al. [27] conducted experimental verification on Zipf's law based on maximum likelihood method. The experimental results show that the slope of Chinese fitting straight line is 1.3 in double logarithmic coordinates, and the slope of content words in this paper is close to that of Chinese fitting, but function words are different. There is a significant difference between the slope of the fitted function words and the content words among various texts, and the difference between them can be seen more intuitively through the fitted curve. The distribution of content words is in a steady downward trend, with the exception of a slight fluctuation in the first few high-frequency words, and the rest of them all decline steadily with a relatively small fluctuation. The first half of the function word curve shows a concave trend, the second half shows a convex trend, and the whole shows a rapid downward trend.

Yu et al. [28] used Zipf's law to fit texts in fifty languages and found that it can be divided into three paragraphs, of which the upper paragraph mainly includes function words, and the middle and lower paragraphs mainly cover content words. Lu et al. [29] studied the word frequency of Chinese, Japanese, and believed that the word frequency distribution of these languages also conforms to Zipf distribution, but the curve fitting index obtained is different from that of Indo-European languages. The authors think that this is because the word language is easier to create new words, and the dictionary space is large, while Chinese language is difficult to create new words, and the dictionary space is limited. As a closed class, function words are difficult to create new words. They mainly play a grammatical role in sentences. Their usages are fixed and their frequency is high. As an open class, content words can constantly create some new words and make the tail of the fitting curve lengthen. The different characteristics of parts of speech just explain the difference in Zipf law distribution.

5. **Conclusion and Future Work.** This article uses the parameters of Zipf's law, the law of word frequency distribution in quantitative linguistics, to explore the difference between content words and function words in modern Chinese. We use news, subtitles, microblog, literature works and online novels five styles of text and make statistics on the parts of speech used in the text. Two fitting methods were used to fit the part-of-speech, the parameters of Zipf's law obtained by fitting and the slope obtained by linear fitting.

According to the comparison, it is found that the parameter of function word is about 1.4, and that of function word is about 2.5, which is a significant difference between them. Since the research in this article does not completely cover all the texts, we must try to verify with more texts if we want to get a more general conclusion. The use of parts of speech varies in different styles, and the deep connection between small parts of speech and style, as well as the determination of some disputed parts of speech by the difference of Zipf's law parameters, are also the direction we want to study next.

Acknowledgment. We are grateful for the comments of all reviewers. This study was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (20YCX154).

REFERENCES

[1] J.Z.Ma, Ma Shi Wen Tong. *The Commercial Press*, 1898.

[2] J.X.Li, New Chinese Grammar. *The Commercial Press*, 1924.

[3] S.S.Ding, The modern Chinese language phrasing talk. *The Commercial Press*, 1961.

[4] S.X.Lu, Issues on Chinese Grammatical Analyses. *The Commercial Press*, 1982.

[5] D.X.Zhu, Lectures on grammar. *The Commercial Press*, 1982.

[6] L.Wang, Modern Chinese Grammar. *The Commercial Press*, 1985.

[7] Y.L.Yuan, The family similarity of the category of parts of speech. *Social Sciences in China,* 1995.

[8] R.Guo, the lexical study of contemporary Chinese. *The Commercial Press,* 2002.

[9] Y.Wang, Y.F.Liu, Q.H.Chen, Zipf distribution of word use in Chinese literature. *Journal of Beijing Normal University(Natural Science),* 2009.

[10] F.Y.He, The applicability of Zipf's law in Chinese language based on words' frequency statistics.*Anhui University.*2011

[11] J. R.Williams,et al., Zipf's law holds for phrases, not words. *Scientific reports.* 2015.

[12] L.Q, Ha., D.Stewart., P.Hanna & F. J.Smith., Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics,* 2006.

[13] N.Hatzigeorgiu, G.Mikros, & G.Carayannis., Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics,* 2001.

[14] B.D.Jayaram., & M.N.Vidya., Zipf's law for Indian languages. *Journal of Quantitative Linguistics*, 2008.

[15] S.W.Choi., Some statistical properties and Zipf's law in Korean text corpus. *Journal of Quantitative Linguistics*, 2000.

[16] G.Dalkilic, & Y.Cebi, Zipf's law and Mandelbrot's constants for Turkish language using Turkish corpus (TurCo). *Lecture Notes in Computer Science*, 2005.

[17] A.Tuzzi., I. I.Popescu, & G.Altmann, Zipf's laws in Italian texts. *Journal of Quantitative Linguistics,* 2009.

[18] R.Xin.et al., Zipf's law and the frequency of Kazak phonemes in word formation. *IOP Conf. Ser.: Mater. Sci. Eng*. 2018.

[19] L.Q.Ha., E.I.Sicilia-Garcia., J.Ming., & F.J.Smith., Extension of Zipf's law to words and character

N-gram for English and Chinese. *Computational Linguistics and Chinese Language Processing,* 2003.

[20] Q.Chen., J.Guo., & Y.Liu., A statistical study on Chinese word and character usage in literatures from the Tang dynasty to the present. *Journal of Quantitative Linguistics,* 2012.

[21] Y.Guan, X.L.Wang, K,Zhang, The Frequency-Rank Relation of Language Units in Modern Chinese Computational Language Model. *Journal of Chinese Information Processing.*1999.

[22] S.T.Piantadosi., Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 2014.

[23] M.P.Marcus., M.A.Marcinkiewicz., & B.Santorini, Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*,1993.

[24] H.T.Liu, W.Huang, Quantitative Linguistics: State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Science),* 2012.

[25] W.J.B.Van Heuven., P.Mandera., E.Keuleers., et al. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 2014.

[26] S.Bernhardsson., L.E.Correa da Rocha., P.Minnhagen., The meta book and sizedependent properties of written language, *New J. Phys.* 2009.

[27] P.Han, G.F.Lu, D.B.Wang, Verification of Zipf's Law Based on Maximum Likelihood Estimation Method .*Information Studies:Theory & Application,* 2012.

[28] S.Y.Yu., et al., Zipf's Law in 50 Languages: Its Structural Pattern, Linguistic Interpretation, and Cognitive Motivation. *ArXiv Preprint ArXiv:1807.01855, 2018.*

[29] L.Lu., Z.K.Zhang., & T.Zhou., Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. *Scientificreports*, 2013.