Qualitative Research on Frequency -Frequency Rank Distribution of Ancient Chinese Characters in Corpus

Haoyang Tian, Ying Liu

Email：honey_tian1991@163.com,yingliu@tsinghua.edu.cn

Lab of Computational Linguistics, School of Humanities, Tsinghua Univ., Beijing 10084, China

Abstract-This article analysis the statistic of ancient Chinese characters' frequency-frequency rank in corpus from Pre-Qin to Qing dynasty, and give the linear curve for the statistic with λ and lnK as parameters, further more, we give a high order function fitting to strengthen the expressive and reduce the complexity of the model, which is superior to the classic Zipf Law curve.

Keywords-frequency-frequencyrank, clustering, installment, Zipf-law

I. INTRODUCTION

GuanYi and other Chinese researchers have been in the study of modern Chinese word - a lot of research on class frequency distribution, research think the modern Chinese word - frequency level distribution accords with Zipf Law. But for China in the ancient Chinese words - frequency level research is relatively small, the distribution of word frequency in the study of ancient Chinese relationship between magnitude and frequency distribution to do scientific reasoning, and does not involve the parameters in the different historical period distribution function changes, and the function fitting effect and considerable error between the observed value. In the ancient Chinese words to study the distribution of the magnitude - frequency have important meaning to the historical evolution of the Chinese language, in the measurement of linguistics history corpora size selection has important reference significance. [1]

In addition to continuous words and names in Chinese place names in ancient China entities such as word, word usually in the form of a single syllable characters involved in the expression, and in China in recent years the development of the theory of "unit" linguistics and popular, confirm the Chinese history from the side in the evolution process of the particularity of language units shown. China's ancient Chinese middle term and classes and ambiguity is very complex, it is not on the ancient Chinese word to realize automatic accurately cut method. Given the prevalence of single words in ancient Chinese and the complexity of cutting word, we are to study in the ancient Chinese words by using the method of measuring magnitude - frequency distribution characteristics, and tries to from different periods of Chinese word - frequency distribution rule in the analysis of the language in the process of historical development shows the characteristics of metering, evolution and the reasons. Researchers in the past to China to study the distribution of the ancient Chinese words - frequency level, through the observation of the large number of real value, and based on Zipf Law assumption of ancient Chinese words to study the distribution of the magnitude - frequency is still a lack of adequate and effective argument.

$$F \propto r^{-\lambda}$$

II. THE ANCIENT CHINESE WORDS - FREQUENCY LEVEL DISTRIBUTION RESEARCH

Zipf Law described in the text word frequency F the nature of the relationship between the magnitude and frequency r:

Zipf Law has a variety of forms, we assume that the frequency of words appears in the table is f, frequency level r, common expression form for the following two kinds:

$$p_r(r) = C_1 r^{\lambda_1} \tag{1}$$

$$p_f(f) = C_2 f^{\lambda_2} \tag{2}$$

We see not hard, above the size of the words in the text of the frequency level R (I) has nothing to do with the size of the text, which in the study of GuanYiDeng people have been confirmed: in the scale above text, words appear PinLv between magnitude and frequency is about the quantitative relationship of euler's constant, has nothing to do with the size of the text. (1) and (2) at the same time also suggests that in the Zipf's law and inverse Zipf's law only needs to know to which a set of parameters, can get parameters of another expression. [2][3]

A. Based on Zipf law assumes that generations of Chinese word - frequency distribution

We based on Zipf Law assumes that level for China in the ancient Chinese words - frequency distribution were studied. We selected the materials according to the division of the pre-qin to the qing dynasty the nine times, in different historical stages of corpora contains all the topics in the text corpus, during this period the corpus from after Taiwan's central institute of ancient Chinese characters, in order to avoid the problem caused by a lack of scale factor data, we from the linguistics of Peking University institute of ancient Chinese corpus to select the appropriate language to carry on the supplement, in addition we went through the students in the direction of ancient Chinese professional artificial verification of proofreading and language version, on the one hand, ensures that the size of the corpus, on the other hand ensure the authenticity and accuracy of corpus. We

through the professional material statistics software Antconc word table, get different period length of word table V parts of the corpus separately (corpus unit: 10 thoundsand words)

TABLE I. We for different stages in the history of corpus to illustrate the theme and scale

	Law	Novel	Poetry	Book	Quotations	Annotation	Screpl	Total	V
Pre-Qin	32.4	——	39.3	124.2	92.4	——	——	290.4	4891
West-Han	66.1	——	69.3	214.3	79.0	12.1	——	440.9	5720
East-Han	79.5	10.1	79.0	235.7	102.7	19.0	——	526.2	5916
Six-Dynasty	22.4	30.7	82.9	116.8	37.2	7.0	——	297.2	6869
Sui-Tang	119.6	97.2	442.9	432.8	105.2	104.4	——	1302.4	7432
Song	226.7	179.2	529.3	449.7	96.2	149.2	4.2	1485.6	7724
Yuan	82.6	34.2	21.7	39.4	21.3	10.2	82.9	282.2	8325
Ming	169.3	3129.7	276.8	320.2	34.2	116.7	10.6	4057.7	8413
Qing	436.1	2197.4	191.3	201.4	12.1	639.7	13.0	3691.2	8492

We assume that the Fr said word frequency of Chinese characters in the table, R is descending order according to the frequency of the frequency of Chinese characters, according to the different periods of Chinese word - frequency level table we trace the Fr - R distribution scatter plot. [4] to study the convenient, we also investigated the exponential variable Fr and R, the distribution of trace lnFr - lnR scatterplot, and linear fitting in SPSS.

In order to verify the Fr - R distribution is in line with the Zipf Law, we assume that the Chinese word frequency in line with the power Law function between magnitude and frequency

$$F_r = K \cdot R^{-\lambda}$$

For the convenience of study, we investigated the F and R take logarithm, about lnF and lnR linear equation is obtained:

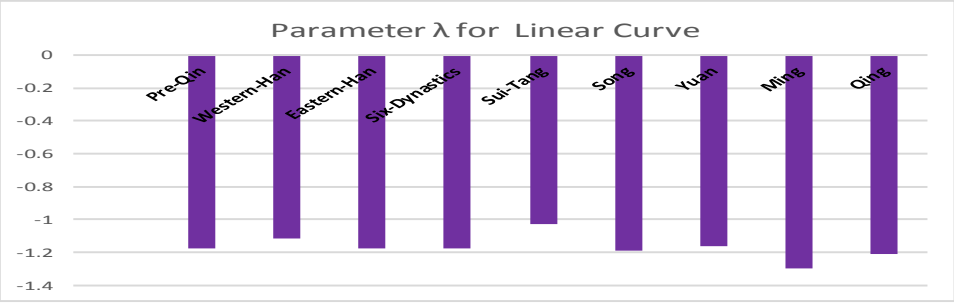
$$\ln F_r + \lambda \ln R = \ln K \qquad (3)$$

That is to say, if we set up our assumptions, variable lnFr and lnR linear correlation.

Tab II is our of different periods of Chinese characters through Matlab lnFr - lnR distribution fitting a linear equation, the value of the parameters is given and the lnK.

TABLE II. Pre-qin - qing lnFr Chinese characters -- -- lnR distribution of linear fitting parameter values

	Pre-Qin	West-Han	East-Han	SixDynasty	Sui-Tang	Song	Yuan	Ming	Qing
λ	-1.1726	-1.1117	-1.1732	-1.1726	-1.0259	-1.1876	-1.1624	-1.2984	-1.2061
lnK	13.2392	12.2859	13.0007	14.1796	12.3482	14.8560	14.2540	15.9995	14.0987



We assume that x = lnR, y = lnFr, lnK = b, so we assume that conforms to the linear relationship between x and y:

$$y = \lambda x + b + \varepsilon \qquad (4)$$

In the function relations, is a random variable, and b for the regression coefficient. According to the principle of least squares, we make:

$$\sum_{i=1}^{N} \varepsilon_i{}^2 = \sum_{i=1}^{N} (y_i - \hat{b} - \hat{\lambda} x_i)^2 = \min_{b,\lambda} \sum_{i=1}^{N} (y_i - \hat{b} - \hat{\lambda} x_i)^2 \quad (5)$$

We will look at value substitution, through to estimate b and N equations and the values and, in the F test, we according to the value of the RSS, ESS, and TSS, and degrees of freedom, and calculated the correlation coefficient of confidence interval. [5] [6]

We assume that $Q = \sum_{i=1}^{N} (y_i - \hat{b} - \hat{\lambda} x_i)^2$, in order to obtain and regression parameters, we make a first order partial derivatives of 0 :

$$\begin{aligned} \frac{\partial (\varepsilon_i^2)}{\partial \lambda} &= -2 \sum_{i=1}^{N} (y_i - \hat{b} - \hat{\lambda} x_i)^2 = 0 \\ \frac{\partial (\varepsilon_i^2)}{\partial \mathrm{b}} &= -2 \sum_{i=1}^{N} x_i (y_i - \hat{b} - \hat{\lambda} x_i)^2 = 0 \end{aligned} \quad (6)$$

Through this group of canonical equation solution, we can easily get linear regression equation:

$$y = \hat{\lambda} x_i + \hat{b} \quad (7)$$

In the process of regression analysis, we can calculate the total sum of squares TSS, explain the sum of squares ESS and residual sum of squares RSS:

To verify lnFr - linear regression equation of lnR is in line with the distribution of the observed value, we F test of the regression equation of the various periods. In Ming dynasty, for example, we assume that H0: linear regression equation was established, TSS, RSS, ESS are respectively x2 variable degrees of freedom, RSS and ESS are independent of each other, then

$$F = \frac{ESS / f_E}{RSS / f_R} = \frac{(N-2)ESS}{RSS} \quad (8)$$

In the F test, statistic, conform to the degrees of freedom F F, significance level and degree of freedom (1841 1), we calculated by look-up table in the Ming dynasty lnFr - lnR linear regression analysis of F distribution is 5081.7, look-up table is F > F alpha, reject H0 assumption, i.e., there is a linear relation between variables x and y.

That is ancient Chinese level of word frequency and frequency distribution accords with Zipf law - law.

B. The word frequency in ancient Chinese cubic function of the magnitude and frequency distribution characteristics

R2 is the SPSS according to least square method of goodness of fit, reflect the fitting function and the observed value (lnFr -- lnR) between the alignment.

We through the same F test, generations of Chinese words from the corpus of magnitude - frequency distribution conforms to the distribution of the three functions, we can see from the scatter diagram of the distribution, high frequency and low frequency field observations and the fitting function has the tendency to deviate from, we can compare the linear fitting and three functions of residual error statistics, you can see the accuracy of the three function fitting. We in Ming dynasty, for example, Fig3.2 are cubic function is linear fitting and fitting residual error statistics:

Therefore, we through the distribution of the three functions of lnFr - lnR fitting. We set y = lnR, x = lnFr, we through the principle of least squares fitting of three functions, we can get all

Parameter	Pre-Qin	West-Han	East-Han	Six-Dynasty	Sui-Tang	Song	Yuan	Ming	Qing
a	-0.0317	-0.0289	-0.0288	-0.0351	-0.0294	-0.0410	-0.0384	-0.0416	-0.0303
b	0.3231	0.2866	0.2512	0.3771	0.2908	0.4613	0.4264	0.4695	0.3066
c	-1.7148	-1.5929	-1.2697	-1.8351	-1.3939	-2.1048	-2.0295	-2.2028	-1.6093
d	11.9996	11.3361	11.0330	12.5097	10.7596	13.0001	12.7710	13.8939	12.3623

C. Based on the stage of high frequency since the period of Chinese studies

We can be found in the past dynasties 1000 words of high frequency they are unique to each period. The high frequency words in the different stage has significant volatility. Different periods of high frequency words use reflects the characteristics of The Times significantly. [8] [9]

Hierarchical clustering based on high-frequency words basically coincide with the periodization of Chinese history research, namely in sui and tang dynasties, for industry, is divided into two sections. In the feature points based on the unique word clustering, can see the sui and tang dynasties, the yuan dynasty and six dynasties as the category, get together for a class of song dynasty, Ming dynasty and qing dynasty, qin, western han dynasty, and get together for a

class of the eastern han dynasty, at the same time, the yuan dynasty, the six dynasties, sui and tang dynasties of the three historical span larger text clustering is a major categories, and the song dynasty, Ming dynasty and qing dynasty, qin, western han dynasty, the eastern han dynasty to another class, Chinese important change period of the six dynasties, sui and tang dynasties and the yuan dynasty, appeared on the expression of Chinese mutation, one of the important performance is a unique character of high frequency words are different. This conclusion and articles of nine historical period before lnFr - lnR distribution of nonlinear fitting of three coefficients have consistency, when Chinese in a larger change, you and the function of three coefficients and adjacent period compared to the value of the difference is obvious.

We through the k - means clustering method, based on the corpus of generations of Chinese top 1000 high-frequency words and top 50 words two unique features for clustering. We can learn by experiment, K - means clustering method of Chinese in the tang dynasty, for line, divided into two parts, with the mainstream of Chinese history studies.

III. CONCLUSION

Based in different periods of large-scale corpus of Chinese word frequency and frequency level, to study the quantitative relation between frequency and frequency level, through mathematical modeling and the method of hypothesis testing for lnFr - lnR curve function fitting and inspection. At the same time analyzed the frequency of high frequency word - level frequency distribution of specific period significant characteristics of Chinese expression. By hierarchical clustering and k means clustering method, with unsupervised method for different periods of text clustering, clearly shows the different period of Chinese text categories and affinity-disaffinity relationship between, with a new method of periodisation of controversial in the history of Chinese thinking and analyzed.

Relative to the level of modern Chinese word - frequency distribution, the ancient Chinese language level of word frequency and frequency distribution between the linear function fitting and observations that there was a considerable error between before and after the article through to the ranking 10% band has carried on the error analysis, think that the ancient Chinese language level of word frequency and frequency distribution of linear fitting function between conform to the overall distribution, but for high frequency and low frequency band level of Chinese word frequency and frequency distribution research have no reference.

Based on three function fitting is analyzed, through the first order and second order partial derivatives were analyzed, and the article surprised to find that different historical period Chinese used in particle concentration on the evolution and mutation, the mutation of the peak and traditional Chinese historical linguistics research conclusions form some kind of corresponding relation. In this paper, through the empirical literature and fitting function partial derivatives of the mathematical logic for the reference, carried on the detailed analysis of the corresponding relations.

REFERENCES

[1] GuanYi, wangxiaolong: "modern Chinese frequency calculation model of language units - frequency level relations," journal of Chinese information 1998.

[2] S.-W. Choi,Some Statistical Properties and Zipf’s Law in Korean Text Corpus,Journal of Quantitative Linguistics,09 Aug 2010.

[3] Hemlata Pande and H.S. Dhami，Model generation for word length frequencies in texts with the application of Zipf’s order approach，Journal of Quantitative Linguistics，2012, Volume 19, Number 4, pp. 249–261.

[4] Damián Zanette and Marcelo Montemurro,Dynamics of Text Generation with Realistic Zipf's Distribution,Journal of Quantitative Linguistics,09 Aug 2010.

[5] B. D. Jayaram and M. N. Vidya,Zipf's Law for Indian Languages,ournal of Quantitative Linguistics,10 Oct 2008.

[6] Arjuna Tuzzi a , Ioan-Iovitz Popescu and Gabriel Altmann，Zipf's Laws in Italian Texts,Journal of Quantitative Linguistics,11 Nov 2009.

[7] Deza, E., & Deza, M. M. (2006). Dictionary of Distances. Amesterdam: Elsevier. Hamming, R. W. (1950). Error detecting and error correcting codes. Bell System Technical Journal, 29(2), 147–160.

[8] Lee, C. Y. (1958). Some properties of nonbinary error-correcting codes. IRE Transactions on, Information Theory, 4(2), 77–82.

[9] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10, 707–710.