On the Applicability of Zipf's Law in Chinese Word Frequency Distribution

Hang Xiao
xiaohang@nus.edu.sg
Department of Chinese Studies, National University of Singapore
Singapore, 117570

Abstract

Zipf's Law uncovers the relationship between word frequency and its rank. This paper addresses applicability of Zipf's Law in Chinese word frequency distribution. The previous studies on Zipf's law in Chinese were primarily based on raw corpus, without word segmentation, hence there are obvious limitations. This study investigates the topic in several large-scale POS-tagged Chinese corpora. The results of these experiments prove that word frequency distribution in Chinese exhibits Zipf's law. The paper further examined the distribution of low frequency word in Chinese corpus, which is estimated by Zipf's law as the majority part of a corpus word list. The result also supports the argument since low frequency words constitute over half of the corpus word occurrences. It indicates that data sparse in statistical approaches could not be magnificently reduced by expanding the corpus scale.

Keywords

Zipf's law; Zipf distribution; word frequency; word frequency distribution; data sparse; Chinese corpus

1. Introduction

Zipf's Law is one of the famous linguistic laws, which describes the statistical distribution of words with different ranks by means of frequency. Zipf's Law uncovers the relationship between word frequency and its position in the word list. The applicability of Zipf's Law in English running text has been examined in large-scale

English corpus such as Brown Corpus. The experiment result suggests that word frequency distribution in English text satisfies Zipf's law (Kucera and Francis, 1967). However, only limited studies have been conducted on the applicability of Zipf's Law in Chinese text. The early studies were not based on well-tagged Chinese corpus, which caused obvious deficiency. This study managed to investigate the applicability of Zipf's Law in Chinese text on the basis of data collected from large-scale tagged Chinese corpus.

In the book "*Human Behaviour and the Principle of Least Effort*", G. K. Zipf (1949) discovered a general principle of human behavior, which was named Principle of Least Effort. The principle of least argues that people will act by way of minimizing their potential average workload. Zipf asserted that the principle of least effort underlies the entire human behavior including language communication. By means of his observation in human languages, Zipf considered that the speaker and hearer manage to minimize their effort: the speaker prefers to use least words (even sentences without punctuation marks) while the hearer expects the speaker to be pellucid, to use more different words, and to express with sufficient language marks; the speaker prefers to use a small vocabulary of common words to reduce their effort, on the other hand, the hearer expects a large vocabulary of rarer words to make the message less ambiguous.

2. Zipf's Law and Word Frequency Distribution

The empirical law uncovered by Zipf reveals some statistical distributions in human language, for instance, the number of the senses of a word form correlates with its frequency of occurrence; a word with higher frequency usually has more morphological or syntactic irregularities. It is examined that the most frequent verbs in a language are also those most likely to be irregular (Tullo and Hurford, 2003). The most famous in these empirical laws is the one on word frequency distribution named Zipf's law. According to the principle of least Effort, people tend to use a small vocabulary to convey rich messages. This tendency leads to the important feature of word frequency distribution in a text, which is only a small number of words have high frequency as well as the majority only have low frequency. Zipf's Law deduces the word frequency distribution that a small number of common words make up the overwhelming majority of word occurrences in the running text, as well as most words in the word list are of low frequency.

The words in a corpus being listed in order of their occurrence frequencies, Zipf's law explores the relationship between the word frequency(f) and its rank(r) in the list by the formula: $f \cdot r = C$. This formula is generally called basic Zipf's formula. Here, r is the rank of a word, f is the word's occurrence frequency, and C is a constant (the value of which depends on the subject under consideration). Zipf estimates that the value of C is approximately 0.1 on the basis of investigation in some English novels.

To further examine the applicability of Zipf's law in different texts, Mandelbrot (1954) extended the study of Zipf's formula. If the word distribution of a corpus is figured on doubly logarithmic axes, Zipf's formula predicts that the graph should be a straight line with slope *-1*. Mandelbrot noticed that Zipf's formula only gives the general shape of words distribution curve and could not well reflect the details. Mandelbrot found that the line is usually not well fit, especially for highest rank words and lowest rank words which are overestimated and underestimated respectively. To achieve a better fit of word frequency distribution described by Zipf's law, Mandelbrot put forward the generalized formula which gives the more general relationship between frequency and rank:

$$f_r = \frac{C}{(r+\alpha)^{\beta}}$$

Here, f_r is words frequency; r is words rank in list; C is a constant; α and β are both constants for the corpus being analyzed. To examine whether the word frequency distribution satisfies Zipf's law, Zipfian curve drawn on doubly logarithmic axes is a general means. If the distribution curve on doubly logarithmic axes is close to a straight line with slope -1, the word frequency distribution in the corpus exhibits Zipf's law.

It is generally known that Zipf's law is an empirical law based on the manual analysis of English text, not a theoretical one. It could not be seen as a common property of a language. However, if it were well approximated, the distributions of many natural or man-made phenomena satisfy Zipf's law, for example, the population distribution of different cities in a country is proved to form a Zipfian curve (Kali, 2003).

3. The Test of Zipf's Law in Running Text

Zipf's law indicates that only a small number of words are frequently used, the majority of words in a corpus are of low frequency. According to Zipf's formula, in a corpus, the word of highest frequency occurs two times as the second highest word; the second highest word occurs two times as the fourth highest word; ...; the frequency of a word ranked *r* is *1/r* times as the highest rank word. For instance, in Brown Corpus, the most frequently used word "*the*" composes 7% of the total word occurrences with 699,971 times per one million word tokens; as described in Zipf's law, the second highest word "*of*" occurs 36,411 times which is nearly the half of the frequency of "*the*"; The 135 most frequently used words occupy the half of total word token frequency. In *The Dictionary of Word Frequency of Contemporary Chinese* (Wang and Chang, 1986), the cumulative frequency of the first 1000 words with highest frequencies is approximately 0.731, which is very close to the value generated by Zipf's law, around 0.748.

3.1 The Prior Estimated Zipf Distribution of Word Frequency

The prior estimation of word frequency distribution could be made based on the basic Zipf's formula: $f = C / r$ ($C=0.1$). Figure 1 is the diagram of the prior estimated Zipf distribution, where $\sum f$ is the cumulative frequency of the word list. In a word list with descending frequency order, cumulative frequency of each word is the sum of frequency value of all words with higher position.

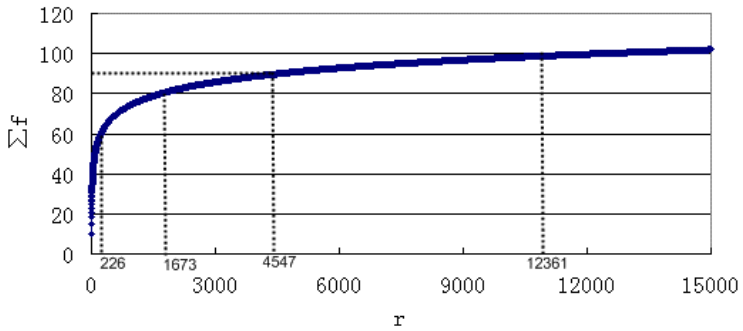

Figure 1. The prior estimated Zipf distribution

Figure 1 shows that a small number of words with high frequency make up the majority of whole word occurrences. The cumulative frequency of the first 30 words is around 30%. This indicates that the 30 most frequently used words constitute 30% of all word occurrences. The cumulative frequency of the 226 high rank words is around 60%; 1673 words 80%, 4547 words 90% and 12361 words approximately 100%. The prior estimated distribution reflects the insight of Zipf's law, only a small number of words make up the main word occurrences of a text. However, noticeably, it is difficult to find a certain text that word occurrences satisfy this estimated distribution. Since natural language is randomized, it is extremely hard to regard the Zipf distribution as a property of a language. Furthermore, the Zipf's law describes only a general pattern of word occurrence; the prior estimated distribution could not be generalized to any specific text or corpus.

3.2 Zipf Distribution in Brown Corpus

Francis and Kucera (1964, 1967) conducted a study to evaluate the word frequency distribution in Brown Corpus. The result of the experiment supports that the word distribution of English corpus exhibits Zipf's law as diagramed in Figure 2.

In Brown Corpus, the most frequently used word "*the*" has a percentage of 7% in whole word occurrence, with 69,971 occurrences per 100,000word tokens. As indicated by Zipf's law, the second highest rank word "*of*" has a percentage of 3.5 %(36,411 occurrences) which is the half of the word "*the*"; the next one is "*and*" with the occurrence of 28,852. The 135 words with highest frequency make up the 50% word occurrence of Brown Corpus.

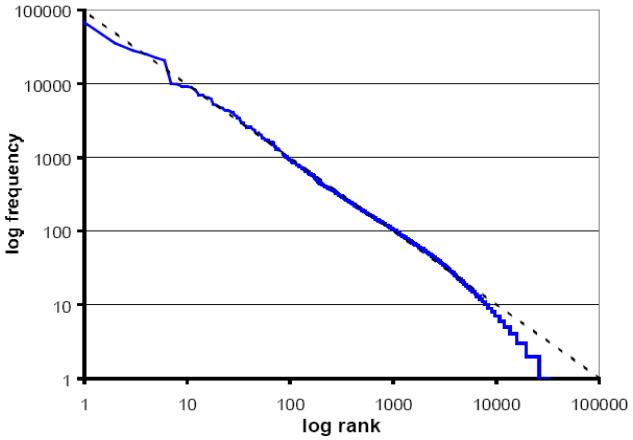

Figure 2. Zipfian curve for Brown Corpus

Figure 2 shows the Zipfian curve for Brown corpus; here, in the doubly logarithmic axes, *r* is the rank of the word; the direct-axis shows the logarithmic word frequency. The Zipfian curve is well fit to the straight line with slope -1. This is a strong evidence to indicate that word frequency distribution in balanced large-scale English corpus strictly follows Zipf's law.

3.3 Zipf Distribution in Other English Corpus

Ha et al. (2003) examined the Zipf's law in "*Wall Street Journal*" Corpus (Paul and Baker，1992). The material of the corpora was selected from the entire text of the journal of 1987, 1988, and 1989. The word counts are 19 million, 16 million and 0.6 million respectively. The result is as figure 3.

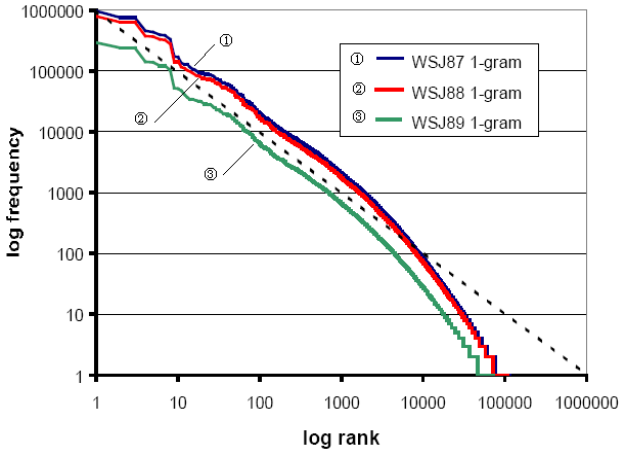

Figure 3. Zipfian curves for WSJ corpora

In figure 3, 1-gram refers to the word in English, while in Chinese corpus, 1-gram refers to the Chinese character. The 1-gram curves for three homogenous corpora are practically parallel with similar slope. It could be another evidence to prove that word occurrence in English satisfies Zipf distribution. In addition, the curves of WSJ87 and WSJ88 which are similar in corpus size are partly overlapped, as well as the small one (WSJ89) has a lower position in the figure. The different size of corpus influences the curve position.

3.4 Zipf Distribution for Chinese N-gram

In Chinese text, 1-gram is known as one Chinese character. However, a Chinese word is made up of uncertain number of characters. As a result, we cannot exactly extract the word list of a Chinese raw corpus without particular word segmentation. In the early studies, N-gram is used to simulate Chinese word units. Ha et al. (2003) conducted a study of word distribution in TREC corpus (Only Chinese materials were used) with the size of 1,954 million Chinese characters. The main content of TREC Chinese corpus includes People Daily newspaper (from 1991.1 to 1993.12) and XINHUA News (from 1994.4 to 1994.9). The TREC Chinese corpus is without word segment and POS tagging. Hence the authors carried out the statistical analysis by means of N-gram. The result is in figure 4.

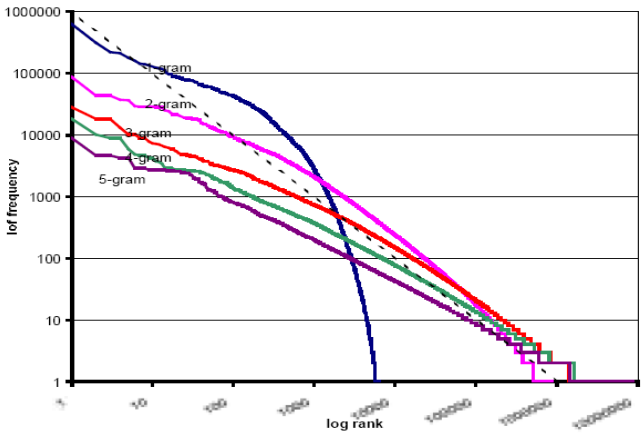

Figure 4. Zipfian curves of N-gram for TREC Chinese corpus

Ha et al. listed part of the most frequently occurred 1-gram, 2-gram and 3-gram extracted from the corpus as follows:

1-gram: 的、国、一、中、在、和、人、了、会、年

2-gram: 中国、发展、经济、星期、国家、企业、经济、人民、记者、社会

3-gram: 新华社、期星期、星期星、百分之、会主义、社会主、社北京、华社北

The combinations extracted by N-gram shows the limitation of the study. It is easy to notice that most of units combined by 3-gram are not exactly Chinese words or phrases. The 2-gram units could strongly indicate that they are from the newspaper-like materials. Moreover, the source of corpus materials seriously influenced the 3-gram extracting. For instance, the fixed-style information of news materials was not been disposed off, so that the combinations, such as "新华社，社北京，华社北", have miraculously high frequencies.

The 1-gram curve in figure 4 supports that the distribution of Chinese characters does not satisfy Zipf's law. This result matches the evaluation of the experiment that we conducted in two Chinese corpora. It could be argued that the distribution of Chinese character in large corpus does not exhibit the Zipfian distribution. This phenomenon may be influenced by the property of Chinese character list that it is generally known as a closed set (20,902 characters in ISO10646). The curves of 2-gram, 3-gram, 4-gram and 5-gram are roughly close to Zipfian distribution.

4. The Applicability of Zipf's Law in Chinese Corpus

Only a small number of studies have been carried out on Zipf's law in Chinese corpus. One obvious reason is the lack of large-scale POS-tagged Chinese corpus. Noticeably, the main corpus used in this study is the large-scale and well-balanced General Contemporary Chinese Corpus (GCCC), developed by the State Language Commission of China. The size of GCCC is around 1 billion Chinese characters with the diachronic materials from 1919 to 2005. The corpus includes 6 categories and 40 sub-categories associated with multi-disciplinary. All the samples in GCCC are well balanced to enhance its representativeness. Most of the materials in GCCC have been well-tagged so that it could be used to conduct a precise word frequency statistics. The minor corpus used in the study is the People Daily Corpus developed by Peking University. Besides, the teaching materials for primary and middle school students from Renmin Educational Press (REP) are used to perform statistic analysis of word use. All these Chinese corpora were segmented and POS-tagged under the similar guideline.

4.1 Zipf Distribution in Large-scale Chinese Corpora

To draw a panorama of Chinese word frequency distribution, we used GCCC to collect statistical data. In the experiment, GCCC was employed in two different ways: a. Entire corpus: 25.5 million words; b. Core corpus: 12.8 million words. The result is shown in Figure 5.

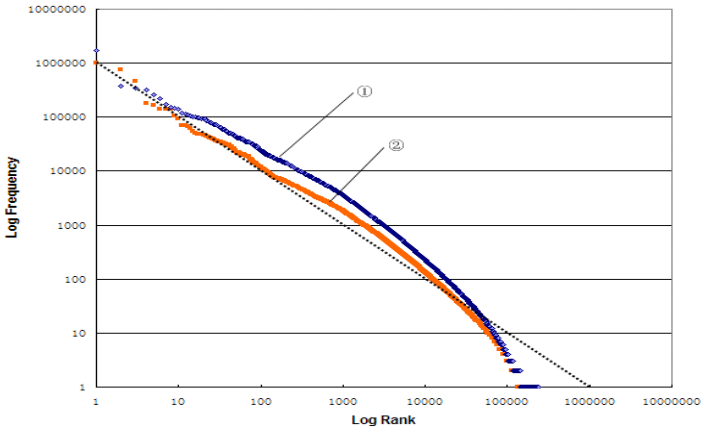

Figure 5. Zipfian curve for GCCC corpus

Here, the Zipf distribution of the entire corpus is shown as curve 1; the core corpus is as curve 2. Obviously, both curves are close to the straight line with slope -1. Hence, it could be argued that the word frequency in large-scale Chinese corpus exhibits Zipf's law.

To make a further evaluation, we constructed 5 small-size corpora by the materials randomly selected from GCCC. The basic information of the corpora is listed in Table 1; the result is diagramed in Figure 6.

Name	Number of		
	Word form	Word occurrence	Character count
CPS-1	41,928	554,501	961,623
CPS-2	42,356	595,320	1,028,893
CPS-3	40,990	544,793	946,070
CPS-4	42,480	558,626	965,647
CPS-5	41,985	551,714	959,492

Table 1. Info of five small-size corpora

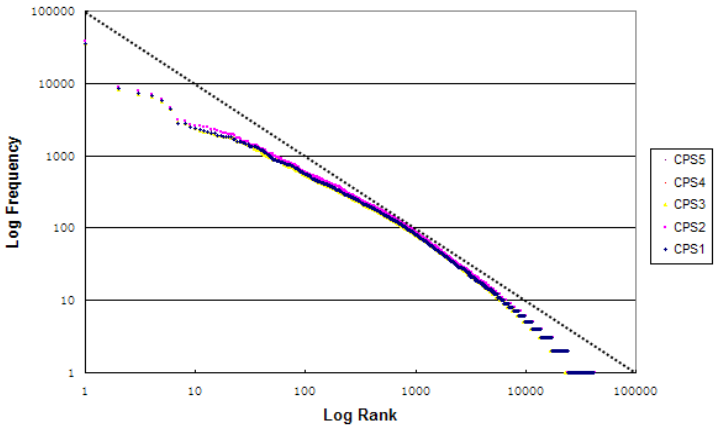

Figure 6. Zipfian curves for 5 small-size corpora

The curves in Figure 6 were highly overlapped, indicating that the five corpora have similar word distributions. The shape of curve supports the conclusion that Chinese word frequency distribution satisfies Zipf's law.

4.2 The Relationship between Corpus Size and Word Distribution

Generally, the number of word tokens is in direct proportion to the size of corpus. Accordingly, a large-scale corpus contains more word tokens as well as is of higher word frequencies. To evaluate the influence of corpus size on word frequency distribution, we examined it in Chinese corpora with different scales. Three different corpora used are: a. teaching materials from textbooks compiled by REP (0.22 million words); b. People Daily newspaper corpus from Peking University (0.92 million words); c. GCCC (25.5 million words). Figure 7 shows the results.

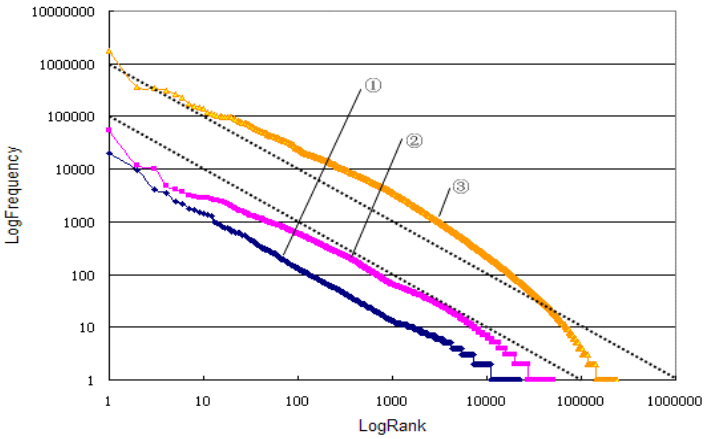

Figure 7. Zipfian curves for different size corpora

It could be concluded from Figure 7 that the size of corpus influences the curve shape of word distribution. The curve of two smaller corpora (curve 1 and 2) are closer to the straight line with slope -1. Curve 1 of 25.5 million words corpus deviates the line slightly. Hence, we argue that if the size of corpus is large enough, word frequency distribution would swerve from Zipf's law gradually, especially the high frequency words and low frequency words. Since larger corpora are hard to obtain, we do not evaluate this assumption.

4.3 The Proportion of Low Frequency Words in Corpus

Zipf's law suggests that only a small number of words are frequently used in human languages, most words are of low frequency. In linguistic studies, the critical reason to build large-scale corpus is to reduce the data sparseness. However, how much data sparse could be reduced by a larger corpus is still unclear. We managed to analyze this feature by the data from corpora of different sizes. In general, there is no exact indicator to judge low frequency. We simply regarded the words fewer than 5 occurrences in corpus as low frequency words. Table 2 shows the proportion of low frequency words in three corpora mentioned above.

Corpus	Size (Million words)	Proportion of low frequency words (%)				
		Occur=1	Occur<=2	Occur<=3	Occur<=4	Occur<=5
Teaching materials	0.22	49.1	64.8	72.6	77.7	81.4
People Daily	0.92	44.9	59.1	66.9	71.9	75.4
GCCC	25.5	40.4	52.1	57.5	61.2	63.9

Table 2. The proportion of low frequency words in three corpora

It is obvious that the proportions of low frequency words in three corpora are significantly high. Firstly, 1-occurrence words account for more than 40% in the corpora; secondly, words fewer than 5 occurrences account for the majority of the corpora; third, the proportion of 1-occurrence words in the largest corpus with 25.5 million words is nearly 40%.

From the data in table 2, we could conclude that the size of corpus influences the distribution of low frequency words. With the growing of corpus size, the proportion of low frequency words declines gradually. Hence, the problem of data sparse is partly solved. On the other hand, we should notice that data sparse could not be completely overcome by increasing the corpus size. As shown in table 1, there are still a large number of low frequency words in large scale corpus.

5. Conclusion

Zipf's law uncovers the possible relationship between word frequency and its rank, and further indicates that the majority of words in corpus are mostly of low frequency. This paper proposed the applicability of Zipf's Law in Chinese word frequency distribution. With the test results illustrated above, the proposition is proved to be fit in Chinese corpus.

Furthermore, according to Zipf's law, we could conclude that data sparse would not be entirely reduced, since a large number of words in human languages are not frequently used. Zipf's law reveals the awkwardness of statistical approaches: almost all words are rare. Therefore, we could have only a small number of words with a great number of examples to analyze. Zipf's law could also benefit corpus linguistics researchers in that the attempt to reduce data sparse by expanding the corpus size is hard to achieve, as half of words in a corpus would continuously be of low frequency. This conclusion maybe answers the question that why few new plans of building large corpus are carried out in these years.

References

Clark, J. L., Lua, K. T. and McCallum, J., 1986, *Using Zipf's law to analyze the rank frequency distribution of elements in Chinese text*, In Proceedings of International Conference on Chinese Computing, pp. 321-324.

Francis, W. and Kucera, H., 1979, *Manual of information to accompany a standard corpus present-day edited American English: for use digital computers*, Providence: Brown University Press.

Gelbukh, A. and Sidorov G., 2001, *Zipf and Heaps laws' coefficients depend on language*, In Proceeding of Conference on Intelligent Text Processing and Computational Linguistics, pp. 332-335

Ha, L.Q., Sicilia-Garcia E.I., Ming J. and Smith F.J., 2003, *Extension of Zipf's law to words and phrases*. Proceedings of the 19th International Conference on Computational Linguistics, pp. 315-320.

Kali R., 2003, *The city as a giant component: a random graph approach to Zipf's law*. Applied Economics Letters, vol. 10, no. 3, pp. 717-720

Kucera, H., and Francis, W., 1967, *Computational analysis of present-day American*

English. Providence: Brown University Press.

Li W., 1992, *Random texts exhibit Zipf's-law-like word frequency distribution,* IEEE Transactions on Information Theory, vol. 38, no. 6, pp. 1942-1845

Mandelbrot B., 1954, *Structure formelledes texts at communication*. Word 10: 1-27.

Manning, C. D. and Schütze H., 1999, *Foundations of statistical natural language processing*, MIT Press.

Miller, G. A., 1965, *Introduction to republication of Zipf (1935)*, Cambridge MA: MIT Press.

Tullo, C., and Hurford, J., 2003, *Modeling Zipfian distributions in language*. In Proceeding of ESSLLI Workshop on Language Evolution and Computation, pp. 62-75.

Wang H., and Chang B. R., 1986, *Dictionary of word frequency of contemporary Chinese*, Beijing: Beijing Language University Press

Zipf G. K., 1935, *The psycho-biology of language: an introduction to dynamic philology*. Houghton Mifflin, Boston.

Zipf G. K., 1949, *Human Behaviour and the Principle of Least-Effort: an introduction to human ecology*. Addison-Wesley, Cambridge, MA.