Bibliography of quantitative studies in Chinese language

Wei HUANG

In the early years of 1980s, Zipf’s and Herdan’s work were introduced into China. Then a few Chinese researchers in linguistics and information science drew attentions to quantitative studies of languages, including theoretical studies of Zipf’s law and its application to frequency distribution of Chinese characters and words.

In recent years, Chinese linguists in increasing numbers are focusing on quantitative linguistics. Both of the theories and the methods of modern quantitative linguistics are comprehensively introduced into China. Meanwhile, Chinese researchers are carrying out quantitative studies in lexicology, syntax, discourse analysis and other branches of linguistics. They have published some quantitative findings of Chinese (Mandarin), English, Russian and other languages.

The following are 32 articles published in Chinese language. Each item consists of 4 parts, the translated bibliographical information in English, the Romanized transliteration according to ISO-7098, the original Chinese information and a brief summary of the article. The bibliography is sorted by years of the publications are in ascending order.


The theory and method of stylo-statistics are introduced from the first part of Herdan’s book Type-token Mathematics: A Textbook of Mathematical Linguistics. It may be the first time when the Chinese researchers made acquaintance with quantitative linguistics.


The derivation and development of Zipf-Mandelbrot law including the work by Estoup, Condon, Zipf, Joos and Mandelbrot are reviewed.


The head part of rank-frequency distribution of Chinese characters in a corpus with size of one million characters can be fitted by the Zipf's law while the tail part approaches exponential distribution. These

1 Beijing Language and Culture University: hwstudio@263.net
findings can be applied in evaluation of optimum code of Chinese characters and evaluation of input scheme.


The word frequency distribution of a Chinese academic article follows Zipf's law.

Cao Congsun (1987). Qipu fu he yuyan de shang. Tianjin shifan daxue xuebao 4, 80-85+73.

Zipf's law and the concept of entropy have been involved in discussion of language evolution. The author claims that Zipf's law shortens the length of language constituent while the entropy enlarges it macroscopically.


The frequency distribution of characters and words in contemporary written Chinese does not obey Zipf's law. The entropy of Chinese words is larger than that of English words.


The word frequency distribution of a Chinese academic article follows Zipf's law.


The self-similarity of Zipfian distribution has been studied experimentally with the fractal method.


The elementary theories and methods of quantitative linguistics, stylo-statistics and mathematical linguistics have been presented. And the early quantitative studies on Chinese, including word frequency and character frequency distribution, stylo-statistics, dialects and speech evolution, have been reviewed.

The exploration in Chinese corpus shows that the frequency distribution of Chinese constituents including characters, words and word bigrams follows Zipf's law. The authors claim that Zipf's law has great effect on many technologies of Chinese automatic processing, especially the construction of Chinese computational language model.


Zipf's law does not fit the whole frequency distribution of Chinese characters. And a method has been presented to describe only the tail of the distribution by using Zipf's law.


Zipf's law and the principle of the least effort, the modified Occam's Razor Principle, the Q-principle and R-principle, the Principle of Relevance have been theoretically discussed.


This article presents investigations of the quantitative lexical characteristics in marine engineering English, including lexical density, zero order, word entropy and perplexity, coverage by the Chinese English Test (CET) Band 4 and CET Band 6 wordlists and the goodness of fit by the Herdan-Heaps model and other models.


The frequency distribution of words and characters in a Chinese lyrics corpus follow Zipf's law. And other experiments on Chinese lyrics in natural language processing including analysis based on time annotation, detecting the repetition of songs identifying rhythms, retrieving songs have been presented.


Based on numerical simulation and regression analysis, this article confirms that Zipf's law is statistically equivalent with the power-law distribution.


The syntactic characteristics of Chinese, including dependent distance and dependent direction, are quantitatively investigated based on 5 dependency treebanks of Chinese. The mean of dependent distance in Chinese is 2.84. And the percentage of the dependency relation of words which are not neighbors is between 40% and 50%. According to the dependency and the direction, Chinese is topologically a SV, VO and AdjN language.


The inter-textual vocabulary repetition of marine English has been examined based on a corpus with size of one million words. According to the Brunet's model of text length and vocabulary size, the inter-textual vocabulary repetition and its 95% confidence interval are calculated. And the 96% of the observed vocabulary repetitions is within the computed 95% confidence interval.


The method of applying the findings in quantitative study of linguistics for scrutinizing text clustering is presented. 16 linguistic structures, which are distributed distinctively between oral and written Chinese, are investigated based on two sample corpora with size of half million words in each.


The analysis of more than 10 thousand Chinese polysemic words extracted from *The Contemporary Chinese Dictionary* shows a strong correlation between polysemy and word frequency.


The inter-textual vocabulary repetition of written English is investigated based on the British National Corpus. And some inspirations in English vocabulary acquisition are discussed.
The word frequency distributions in Chinese literature, including *Dream of Red Chamber*, *Selected Works of Mao Tse-tung* and *Selected Works of Deng Xiaoping*, follow Zipf's law.


The investigation in technical English text based on the Jiao Da English for Science and Technology corpus confirms that both the relationship between compounding propensity and stem frequency, and the relationship between compounding propensity and length of stem follow \( y = ax^b \), and that the relationship between compounding propensity and polylexy follows \( y = a + bx^2 \).


The word frequency distribution in a Kazak corpus with size of 300 thousand words follows Zipf's law.


The experiment of simulating random data series following Zipf-Mandelbrot’s law reveals that the parameter \( \rho \) in the formula \( S \propto (r + \rho)^{-a} \) reflects saturation effect of sampled data.

Xia Huiyan, Sun Fenglan (2011). Quantitative relationship between word length and polysemy in English. *Foreign Languages and Their Teaching* 3, 44-49.

The investigation in the British National Corpus shows that the relationship between polysemy and word length follows the power model \( y = Ax^{-b} \), while the relationship between synonymy and word length does not follow the model \( y = Ax^{b}e^{cx} \), but the model \( y = a + bx + cx^2 + dx^3 \) does well.


The comparison of Ordinary Least Square and Maximum Likelihood Estimation in fitting of the
distribution of Zipf's law with 6 corpora, including 3 Chinese ones and 3 English ones, shows that the Maximum Likelihood Estimation is much better.


The relationship of quantitative linguistics and mathematical linguistics is discussed, and Zipf's law, the Menterath-Altman law, the Fucks-Čebanov law and the Piotrowski-Altman law are reviewed.


This article reviews the history and development of quantitative linguistics, introduces the classical theories and achievements, and discusses the methodology of quantitative linguistics. It is the first time that the quantitative linguistics and synergetic linguistics are comprehensively introduced into Chinese.

**Dong Hong (2012).** Inter-textual vocabulary growth patterns for Shakespearean drama. *Journal of Qiqihar University (Philosophy and Social Science) 2,* 125-127.


The investigation in the Shakespearean Play Corpus shows that the inter-textual vocabulary growth of the hapax legomena follows the Brunet's model $V = a (\ln N)^b$.


The investigation in the Lancaster Corpus of Mandarin Chinese and the Spoken Corpus of Mandarin Chinese shows that the relationship between word length and word frequency in Chinese texts follows the power law $y = ax^b$, and the parameter $a$ is distinguishable between oral speech and written texts.

**Wang Yong, Liu Haitao (2013).** Quantitative properties of Russian nouns. *Journal of Zhejiang University (Humanities and Social Sciences) 43*(6), 174-186.


The quantitative investigation in a Russian corpus shows that 1) the Russian is a dependent-final language, 2) the major types of nominal structures in Russian are concordant attribute and non-concordant attribute relations, and 3) the word orders in the Russian nominal structures tend to follow certain patterns.

**Zheng Chen, Hu Manfeng (2013).** Zipf's law in words and characters in Moyan's literature. *Journal of*
The word and the character frequency distribution in 3 novels by Mo Yan, who is a Nobel Prize winner for literature, follow Zipf's law.