Glottometrics 31, 2015, 76-84

Bibliography of quantitative studies in Chinese language

Wei HUANG[1]

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.

Xiao Shensheng (1982). G. Herdan's stylo-statistics. *Language Study 2, 104-117.*
Xiao Shensheng (1982). G. Herdan de yanyu fengge tongjixue. Yuyan yanjiu *2, 104-117.*
萧申生. (1982). G.Herdan 的言语风格统计学. 语言研究 *2, 104-117.*

The theory and method of stylo-staticstics 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.

Feng Zhiwei (1983). The origin and development of Zipf's law. *Information Science 4(2), 37-42*.
Feng Zhiwei (1983). Qipufu dinglü de lailongqumai. *Qingbao kexue 4(2), 37-42.*
冯志伟. (1983). 齐普夫定律的来龙去脉. 情报科学 *4(2), 37-42.*

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

Shi Guiqing, Xu Bingzheng (1984). On the frequency distribution, optimum coding and input scheme for Chinese characters. *ACTA Electronica Sinica 12(4), 94-96.*
Shi Guiqing, Xu Bingzheng (1984). Hanzi zipin fenbu、zuijia bianma yu shuru wenti. *Dianzi xuebao 12(4), 94-96.*
石贵青、徐秉铮. (1984). 汉字字频分布、最佳编码与输入问题. 电子学报 *12(4), 94-96.*

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.

Xu Wenxia (1986). Zipf's law and word frequency distribution of Chinese. *Information Science 7(1), 29-36.*

Xu Wenxia (1986). Qipufu dinglü yu zhongwen cipin fenbu jili. *Qingbao kexue 7(1), 29-36.*

许文霞. (1986). 齐普夫定律与中文词频分布机理. 情报科学 *7(1), 29-36*.

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

Cao Congsun (1987). Zipf's law and entropy of languages. *Journal of Tianjin Normal University 4, 80-85+73.*

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

曹聪孙. (1987). 齐夫定律和语言的熵. 天津师范大学学报 *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.

Wang Dejin (1988). The probability distribution and entropy in printed Chinese. *Journal of Beijing Institute of Aeronauties and Astronautics 4, 89-94.*

Wang Dejin (1988). Hanyu zi、ci de gailü fenbu he yi jie shang de yanjiu. *Beijing hangkong xueyuan xuebao 4, 89-94.*

王德进. (1988). 汉语字、词的概率分布和一阶熵的研究. 北京航空学院学报 *4, 89-94.*

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.

Wang Chongde, Lai Ling (1989). The Chinese collected work of Zipf's distribution. *Information Science 10(2), 1-8+42+79.*

Wang Chongde, Lai Ling (1989). Hanyu wenji de Qifu fenbu. *Qingbao kexue 10(2), 1-8+42+79.*

王崇德、来玲. (1989). 汉语文集的齐夫分布. 情报科学 *10(2), 1-8+42+79.*

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

Zhao Laiyuan (1996). Fractal representation of Zipf's law. *Journal of the China Society for Scientific and Technical Information 15(4), 74-79.*

Zhao Laiyuan (1996). Qifu dinglü de fenxing tixian. *Qingbao xuebao 15(4), 74-79.*

赵来远. (1996). 齐夫定律的分形体现. 情报学报 *15(4), 74-79.*

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

Chen Hailun (1996). A brief introduction to quantitative linguistics. *Journal of Yulin Teachers College (Philosophy and Socical Science) 17(1), 37-41+56.*

Chen Hailun (1996). Jiliang yuyanxue shuo lüe. *Yulin shizhuan xuebao 17(1), 37-41+56.*

陈海伦. (1996). 计量语言学说略. 玉林师专学报 *17(1), 37-41+56.*

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.

Guan Yi, Wang Xiaolong, Zhang Kai (1999).The frequency-rank relation of language units in modern Chinese computational language model. *Journal of Chinese Information Processing 13(2), 9-16.*

Guan Yi, Wang Xiaolong, Zhang Kai (1999). Xiandai hanyu jisuan yuyan moxing zhong yuyan danwei de pindu - pinji guanxi. *Zhongwen xinxi xuebao 13(2), 9-16.*

关毅、王晓龙、张凯. (1999). 现代汉语计算语言模型中语言单位的频度-频级关系. 中文信息学报 *13(2), 9-16.*

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.

You Rongyan (2000). Zipf's law and the distribution of Chinese character frequency. *Journal of Chinese Information Processing 14(3), 60-65.*

You Rongyan (2000). Zipf dinglü yu hanzi zipin fenbu. *Zhongwen xinxi xuebao 14(3), 60-65.*

游荣彦. (2000). Zipf 定律与汉字字频分布. 中文信息学报 *14(3), 60-65.*

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.

Jiang Wangqi (2005). Zipf and the principle of the least effort. *Journal of Tongji University (Social Science Section) 16 (1), 87-95.*

Jiang Wangqi (2005). Zipf yu shengli yuanze. *Tongji daxue xuebao (shehui kexue ban) 16(1), 87-95.*

姜望琪. (2005). Zipf 与省力原则. 同济大学学报(社会科学版) *16(1), 87-95.*

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.

Fan Fengxiang (2006). Quantitative lexical description of marine engineering English. *Journal of Dalian Maritime University (Social Sciences Edition) 5(3), 161-164.*

Fan Fengxiang (2006). Lunji yingyu cihui de lianghua tezheng. *Dalian haishi daxue xuebao (shehui kexue ban) 5(3), 161-164.*

范凤祥. (2006). 轮机英语词汇的量化特征. 大连海事大学学报(社会科学版) *5(3), 161-164.*

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.

Zheng Yabin, Liu Zhiyuan, Sun Maosong (2007). Statistical features of Chinese song lyrics and its application to retrieval. *Journal of Chinese Information Processing 21(5), 61-67.*

Zheng Yabin, Liu Zhiyuan, Sun Maosong (2007). Zhongwen geci de tongji tezheng ji qijiansuo yingyong. *Zhongwen xinxi xuebao 21(5), 61-67.*

郑亚斌、刘知远、孙茂松. (2007). 中文歌词的统计特征及其检索应用. 中文信息学报 *21(5), 61-67.*

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.

Gong Xiaoqing, Wang Zhan (2008). A note on Zipf's law. *Complex Systems and Complexity Science 5(3), 73-78.*

Gong Xiaoqing, Wang Zhan (2008). Guanyu Zipf lü de yidian zhuji. *Fuza xitong yu fuzaxing kexue 5(3), 73-78.*

龚小庆、王展. (2008). 关于 Zipf 律的一点注记. 复杂系统与复杂性科学 *5(3), 73-78.*

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

Liu Haitao (2008). Quantitative study of Chinese grammar based on dependency treebank. *Yangtze River Academic 3, 120-128.*

Liu Haitao (2008). Jiyu yicun shuku de hanyu jufa jiliang yanjiu. *Changjiang xueshu 3, 120-128.*

刘海涛. (2008). 基于依存树库的汉语句法计量研究. 长江学术 *3, 120-128.*

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.

Fan Fengxiang (2008). Inter-textual vocabulary repetition of marine engineering English. *Journal of Dalian Maritime University (Social Sciences Edition) 7(2). 128-132.*

Fan Fengxiang (2008). Lunji yingyu de pianji cihui chongfulü. *Dalian haishi daxue xuebao (shehui kexue ban) 7(2), 128-132.*

范凤祥. (2008). 轮机英语的篇际词汇重复率. 大连海事大学学报(社会科学版) *7(2), 128-132.*

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.

Huang Wei, Liu Haitao (2009). Application of quantitative characteristics of Chinese genres in text clustering. *Computer Engineering and Applications 45(29), 25-27.*

Huang Wei, Liu Haitao (2009). Hanyu yuti de jiliang tezheng zai wenben julie zhong de yingyong. *Jisuanji gongcheng yu yingyong 45(29), 25-27.*

黄伟、刘海涛. (2009). 汉语语体的计量特征在文本聚类中的应用. 计算机工程与应用 *45(29), 25-27.*

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.

Wang Hui (2009). Polysemy: meaning, length and frequency. *Chinese Language 2, 120-130+191.*

Wang Hui (2009). Ciyi· cichang· cipin --《Xiandai hanyu cidian》(di 5 ban) duoyi ci jiliang fenxi. *Zhongguo yuwen 2, 120-130+191.*

王惠. (2009). 词义· 词长· 词频—《现代汉语词典》(第 5 版)多义词计量分析. 中国语文 *2, 120-130+191.*

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

Luo Weihua, Deng Yaochen (2009). Pattern of inter-textual vocabulary repetition in English based on BNC. *Foreign Language Teaching and Research 41(3), 224-229.*

Luo Weihua, Deng Yaochen (2009). Jiyu BNC yuliaoku de yingyu pianji cihui chongfu moshi yanjiu. *Waiyu jiaoxue yu yanjiu 41(3), 224-229.*

罗卫华、邓耀臣. (2009). 基于 BNC 语料库的英语篇际词汇重复模式研究. 外语教学与研究 *41(3), 224-229.*

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.

Wang Yang, Liu Yufan, Chen Qinghua (2009). Zipf distribution of words used in Chinese literature. *Journal of Beijing Normal University (Natural Science) 45(4), 424-427.*

Wang Yang, Liu Yufan, Chen Qinghua (2009). Hanyuyan wenxue zuopin zhong cipin de Zipf fenbu. *Beijing*

shifan daxue xuebao (ziran kexue ban) 45(4), 424-427.

王洋、刘宇凡、陈清华. (2009). 汉语言文学作品中词频的 Zipf 分布. 北京师范大学学报(自然科学版) *45(4), 424-427.*

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.

Deng Yaochen, Li Bingbing (2010). Study of English compounding propensity in synergistic linguistics. *Foreign Languages and Their Teaching 4, 19-24.*

Deng Yaochen, Li Bingbing (2010). Yingyu fuheci shengcheng quxiang de xietong yuyanxue yanjiu. *Waiyu yu waiyu jiaoxue 4, 19-24.*

邓耀臣、李冰冰. (2010). 英语复合词生成趋向的协同语言学研究. 外语与外语教学 *4, 19-24.*

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 $y = a + bx^2$.

Wang Hua, Gulila Altenbek (2010).A corpus–based study on frequency statistics of Kazak words. *Computer Engineering 36(24), 59-61.*

Wang Hua, Gulila Altenbek (2010). Jiyu yuliao de hasakeyu cipin tongji yanjiu. *Jisuanji gongcheng, 36(24), 59-61.*

王花、古丽拉· 阿东别克. (2010). 基于语料的哈萨克语词频统计研究. 计算机工程 *36(24), 59-61*.

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

Tang Lian, Wang Dahui (2011). An explanation of shift parameter ρ in Zipf-Mandelbrot's law. *Journal of Beijing Normal University (Natural Science) 47(1), 97-100.*

Tang Lian, Wang Dahui (2011). Guanyu Zipf-Mandelbrot lü zhong canshu ρ de yizhong jieshi. *Beijing shifan daxue xuebao (ziran kexue ban) 47(1), 97-100.*

唐莲、王大辉. (2011). 关于 Zipf-Mandelbrot 律中参数 ρ 的一种解释. 北京师范大学学报(自然科学版) *47(1), 97-100.*

The experiment of simulating random data series following Zipf-Mandelbrot's law reveals that the parameter ρ 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.*

Xia Huiyan, Sun Fenglan (2011). Yingyu cihui changdu yu cihui yiyi guanxi de jiliang yuyanxue yanjiu. *Waiyu yu waiyu jiaoxue 3, 44-49.*

夏慧言、孙凤兰. (2011). 英语词汇长度与词汇意义关系的计量语言学研究. 外语与外语教学 *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.

Lu Gaofei, Han Pu, Shen Si (2012). Comparative empirical study on Zipf's law with two fitting methods. *Library and Information Service 56(24), 71-76+126.*

Lu Gaofei, Han Pu, Shen Si (2012). Liang zhong Zipf dinglü nihe fangfa de duibi shizheng yanjiu. *Tushu qingbao gongzuo 56(24), 71-76+126.*

路高飞、韩普、沈思. (2012). 两种 Zipf 定律拟合方法的对比实证研究. 图书情报工作 *56(24), 71-76+126.*

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.

Feng Zhiwei (2012). Studying language by quantitative method. *Foreign Language Teaching and Research 44(2), 256-269+321.*

Feng Zhiwei (2012). Yong jiliang fangfa yanjiu yuyan. *Waiyu jiaoxue yu yanjiu 44(2), 256-269+321.*

冯志伟. (2012). 用计量方法研究语言. 外语教学与研究 *44(2), 256-269+321.*

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

Liu Haitao, Huang Wei (2012). Quantitative linguistics: state of the art, theories and methods. *Journal of Zhejiang University (Humanities and Social Sciences) 42(2), 178-192.*

Liu Haitao, Huang Wei (2012). Jiliang yuyanxue de xianzhuang、lilun yu fangfa. *Zhejiang daxue xuebao (renwen shehui kexue ban) 42(2), 178-192.*

刘海涛、黄伟. (2012). 计量语言学的现状、理论与方法. 浙江大学学报(人文社会科学版) *42(2), 178-192.*

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.*

Dong Hong (2012). Shashibiya xiju de pianji cihui zengzhang moshi. *Qiqihar daxue xuebao (zhexue shehui kexue ban) 2, 125-127.*

董红. (2012). 莎士比亚戏剧的篇际词汇增长模式. 齐齐哈尔大学学报(哲学社会科学版) *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 $V = a\ (ln\ N)^b$.

Deng Yaochen, Feng Zhiwei (2013). A quantitative linguistic study on the relationship between word length and word frequency. *Journal of Foreign Lanugages 36(3), 29-39.*

Deng Yaochen, Feng Zhiwei (2013). Cihui changdu yu cihui pinshuo guanxi de jiliang yuyanxue yanjiu. *Waiguoyu (Shanghai waiguoyu daxue xuebao) 36(3), 29-39.*

邓耀臣、冯志伟. (2013). 词汇长度与词汇频数关系的计量语言学研究. 外国语(上海外国语大学学报) *36(3), 29-39.*

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.*

Wang Yong, Liu Haitao (2013). Eyu mingci de jiliang tezheng yanjiu. *Zhejiang daxue xuebao (renwen shehui kexue ban) 43(6), 174-186.*

王永、刘海涛. (2013). 俄语名词的计量特征研究. 浙江大学学报(人文社会科学版) *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*

Jiangnan University (Natural Science Edition) 12(3), 347-350.

Zheng Chen, Hu Manfeng (2013). Moyan zuopin zhong zipin, cipin de Zipf fenbu. *Jiangnan daxue xuebao (ziran kexue ban) 12(3), 347-350.*

郑晨、胡满峰. (2013). 莫言作品中字频、词频的 Zipf 分布. 江南大学学报(自然科学版) *12(3), 347-350.* 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.