Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 358 (2005) 545–550
PHYSICA A
www.elsevier.com/locate/physa

True reason for Zipf's law in language

Wang Dahui[a,*], Li Menghui[b], Di Zengru[b]

[a]Department of Systems Science, School of Management, Beijing Normal University, Beijing 100875, China
[b]Center for Complexity Research, Beijing Normal University, Beijing 100875, China

Received 1 April 2005
Available online 2 June 2005

Abstract

Analysis of word frequency have historically used data that included English, French, or other language, data typically described by Zipf's law. Using data on traditional and modern Chinese literatures, we show here that Chinese character frequency stroked Zipf's law based on literature before Qin dynasty; however, it departed from Zipf's law based on literature after Qin dynasty. Combined with data about English dictionaries and Chinese dictionaries, we show that the true reason for Zipf's Law in language is that growth and preferential selection mechanism of word or character in given language.

Keywords: Zipf's law; Language; Mechanism

1. Introduction

Many languages, such as English, French, Spanish, have been found to exhibit some universal characteristics called Zipf's law [1–3],which read as $p(r) \propto r^{-\alpha}$, where r is the rank of particular word, $p(r)$ is the frequency of the rth word occurrence in the text, and α is an exponent. This law describes surprisingly diverse natural and social phenomena, including percolation processes [4], immune system response [5], city sizes [1,6], and aspects of Internet traffic [7,8]. There are some explanations for this law in language. For example, Zipf himself proposed "least effort of human

*Corresponding author. Tel.: +86 10 58807876; fax: +86 10 58807876.
E-mail address: wangdh@bnu.edu.cn (W. Dahui).

0378-4371/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.04.021

behavior principal'' [1,2]. H.Simon had advanced random selection mechanism for power law generating [9]. And some other authors looked at the processes of text generating as a Markov processes [3], or self-organized criticality [10,11], as well as other explanations [12–14]. Since these explanations conform to empirical results based on English or French corpus, could we say that they are comprehensive explanations for Zipf's law in language? In this paper, our answer is no. Here we present the empirical results about Chinese literature, which are quite different from results about other languages. We also set up a dynamic model to simulate the process of text generating and explain Zipf's law in a variety of languages. Our conclusion about true reason for Zipf's law in language is that growth and preferential selection of word or character mechanism in given language.

2. Empirical results about Chinese character

In this paper,we investigate the Chinese literature extensively [15], including: excavated oracles, characters inscribed on bones or tortoise shells in the Shang Dynasty (16th–11th century B.C.); excavated bronze inscription, of the late Shang Dynasty (14th–11th century B.C.) and the Zhou Dynasty (11th–6th century B.C.); classic poetry, of 11th-6th century B.C.; Erya, the first dictionary of Chinese about 4th century B.C.; Shi Ji, of Han Dynasty (100 B.C.); poetry, of Tang Dynasty (618–904A.D.); Ci,of the Song Dynasty (960–1279); the Dream of Red Chamber, of Qing Dynasty (1750s); Collection of Chairman Mao; the latest novels Mo Fa Xue Tu and Liang Jian posted on the internet. The empirical results are shown in Fig. 1.

The empirical results of Chinese are quite different from those of English. The empirical results of earlier are different from those of later Chinese. The earlier Chinese obey Zipf's law, but the later obviously depart from Zipf's law. What causes this difference between Chinese and other languages? Let us pay attention to the some features of Chinese characters and English words. Before the Qin dynasty, Chinese characters were in infancy and different in various areas of China. After Emperor Qin Shihuang unified the characters, the Chinese language became mature. It is difficult to create new characters because Chinese characters are pictographs, and the number of Chinese characters has grown very slowly, from 10 000 to 50 000 over last 2000 years. So, the available number of Chinese characters for any author is almost fixed. On the other hand, the words of other language, such as English, new words are introduced constantly and the number of words grows very fast compared with Chinese character. The available words for authors are unlimited. To illustrate this difference, we plotted the number of the characters in Chinese dictionaries and those of English dictionaries in Fig. 2(a) [16]. In addition to the growth ratio of Chinese characters or English words over time, the growth ratio of the number of different individual characters or words used by an author as the text length increases is also quite different between Chinese and English. This observation is shown in Fig. 2(b).

As Zipf's law dose not apply under the condition that Chinese characters size does not expand, could we say that the character or word growth is important ingredient for Zipf's law in language? Considering that we can look at all words and characters

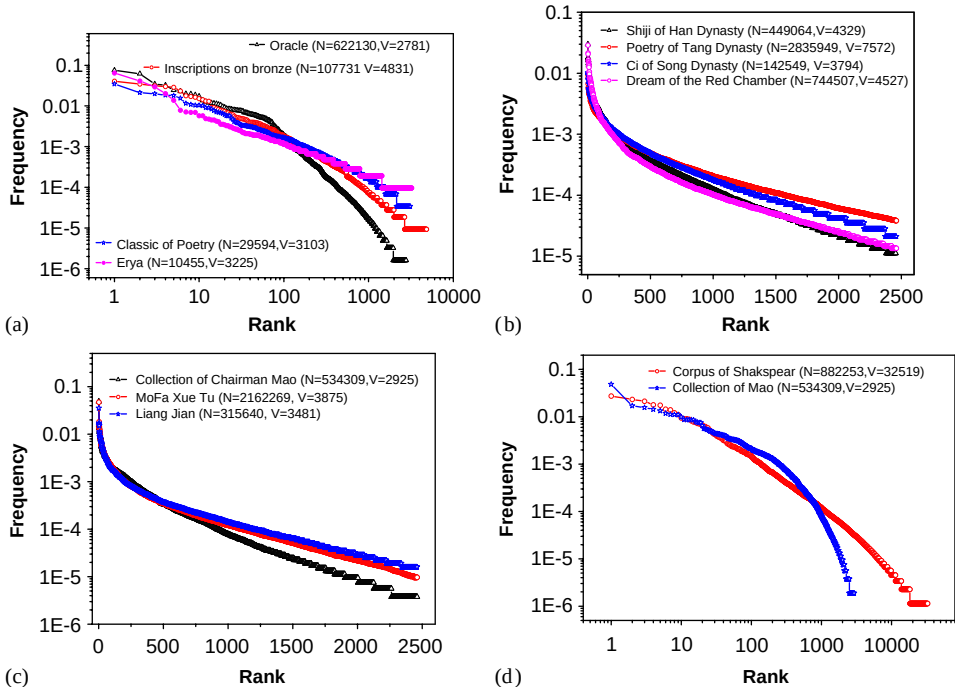


Fig. 1. Empirical results about Chinese. Where N is length of the text, and V is the number of different characters. (a) These literatures pre-date emperor Qin Shihuang's unification of Chinese characters. Although the frequency to rank is not exactly the same, these earliest literatures conform to Zipf's law. (b) Ancient literature later than emperor Qin Shihuang do not conform to Zipf's law. (c) Frequency to rank results of modern literature don't strike Zipf's law. (d) Frequency to rank statistics of Shakespeare's corpus and that of the Collection of Chairman Mao are shown to illustrate the difference between English and modern Chinese.

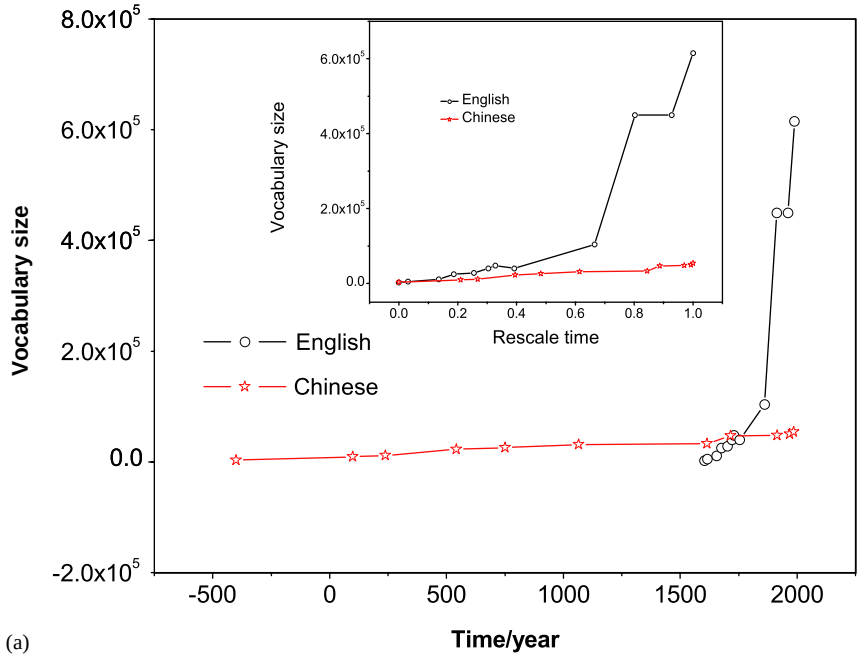
as only some kind of different symbols without any meaning and look at the text as series of these symbols, the answer is positive. In fact, there are more than 100 natural languages in which the size of the alphabet ranges from 14 to 60 that obey the following law, which is different from Zipf's law (3):

$$p(k) = A - D\ln(k)\,, \tag{1}$$

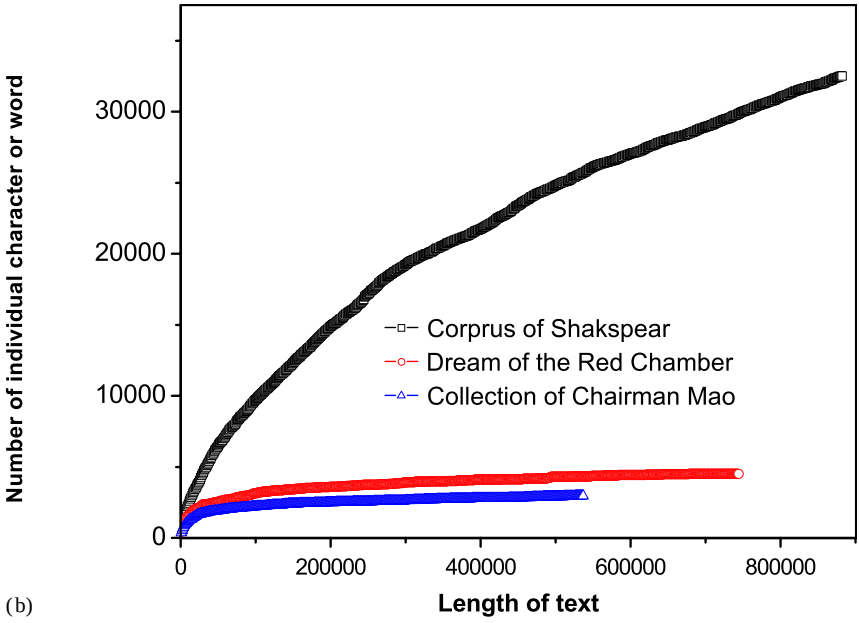
where $p(k)$ is frequency of the kth alphabet, A and D are constants. It is obviously that the size of an alphabet in a language is almost constant.

3. Dynamic model and simulation results

From above observations, we present a dynamic model that explains the empirical behaviors of Chinese characters and other language words as result of two interacting processes: firstly, characters or words added to the text with a global



(a)



(b)

Fig. 2. Vocabulary size expansion and the number of individual characters/words in text. (a) The figure shows the number of individual entries/characters collected in the dictionary over dictionary published ages. The axis of the inner graph is rescaled. It is obviously that English vocabulary expands much faster than that of Chinese. (b) This figure shows the number of individual words/characters over the length of the text. The total number of different characters is generated very quickly in Chinese, whereas total word usage grows as text length increases in English.

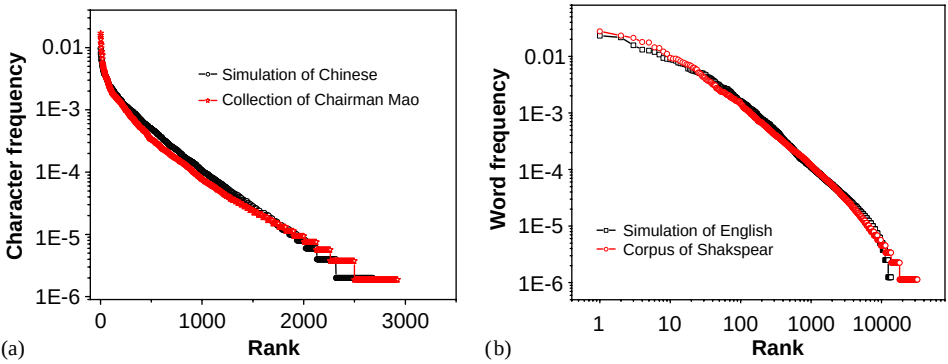


Fig. 3. Simulation results. (a) Simulation about the Chinese case, where $V_0 = 40\,000, \sigma = 100, \beta = 0$. The simulation step is 500 000. The simulation result conforms to the empirical result about the Collection of Chairman Mao. (b) Simulation in the English case, where $V_0 = 5000, \sigma = 60, \beta = 8$. The simulation result conforms to the empirical results about Shakespeare's corpus.

memory effect reflected as characters or words preferential selection; secondly, with the growth of possible characters or words that could be selected, the size of the vocabulary expands. Suppose that the initial vocabulary size is V_0 and increased by β at each step,

$$V(t+1) = V(t) + \beta\,. \tag{2}$$

The situation where β equals zero represents the case of Chinese. At each step t, one character or word is added to text beginning with just one word at $t = 1$ such that, at any step, the length of the text is t and the ith character or word occurrences $n(i, t)$ times in the text. Considering that the global memory effect and preferential select, the more a character or word is used, the more likely it is to be used again, we give the probability of ith character or word is added at $t + 1$ as following:

$$p(i,t) = \frac{1 + \sigma n(i,t)}{V_0 + \sigma t + \beta t}, \quad i = 1, 2, \ldots V(t-1), V(t)\,, \tag{3}$$

where σ indicates strength of global memory effect and preferential selection. From this model, we can determine the relationship between the frequency of ith character/word and its rank. The results are shown in Fig. 3.

4. Summary

In closing, we investigate the character frequency of traditional and modern Chinese corpus and find that Chinese character frequency in traditional corpus before Qin dynasty obey Zipf's law but that in Chinese corpus after Qin dynasty does not obey Zipf's law. We also set up a dynamic model to simulate the text generation and the results of our model shown in Fig. 3 are consistent with Chinese and other language, so we recognize that the mechanism of word/character growth and

preferential selection of our model is a real reason for Zipf's law in language. From this model, we can give a conjecture about Chinese language as following: although almost no new characters are introduced into Chinese, new phrases are continuously into Chinese for explanation for emergent novelty. So, we can conjecture that frequency to rank statistics of Chinese phrase may conform to Zipf's law.

Acknowledgements

This research was supported in part by the National Science Foundation of China under Grant No. 70471080, 70371072. The authors thank Prof. Hu C.K. for his helpful discussion.

References

[1] G.K. Zipf, Human Behavior and the Principal of Least Effort, Addison-Wesley, Cambridge, MA, 1949.
[2] Romon Ferrer i Cancho, R.V. Solé, PNAS 100 (2002) 788–791.
[3] I. Kanter, D.A. Kessler, Phys. Rev. Lett. 74 (1995) 4559.
[4] M.S. Watanabe, Phys. Rev. E 53 (1996) 4187.
[5] J.D. Burgos, P. Moreno-Tovar, Biosystems 39 (1996) 227.
[6] L. Breslau, et al., Proceedings of INFOCOM '99, 21–25 March 1999, New York, IEEE Press, Piscataway, NJ, 1 (1999) 126–134.
[7] J.J. Ramsden, G. Kiss-Haypal, Physica A 277 (2000) 220.
[8] R. Albert, A.L. Barabsi, Rev. Mod. Phys. 74 (2002) 47–97.
[9] H. Simon, Biometrika 42 (1955) 425–440.
[10] P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. 59 (1987) 381.
[11] P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. 60 (1988) 2347.
[12] W. Li, IEEE Trans. Inform. Theory 38 (1992) 1842.
[13] Rommon Ferrer i Cancho, R.V. Solé, Adv. Complex Syst. 5 (2002) 1–6.
[14] B. Mandelbrot, The Fractal Structure of Nature, New York, Freeman, 1983.
[15] The frequency of character in oracles, classic of poetry, Erya, and Shiji come from Chinese ancient texts center of the Chinese University of Hong Kong (available at http://www.chant.org). The frequency of character of bronze inscription is from Index of Compilation of Yin and Zhou Bronze Inscriptions (Zhonghua book company, 2001). We count the character frequency of other literature except for above mentioned. The corpus of following materials are draw from the internet: Poetry of Tang Dynasty, Ci of Song Dynasty, Dream of Red Chamber, Collection of Chairman Mao, the latest novels Mo Fa Xue Tu and Liang Jian.
[16] The English dictionaries including: A Table Alphabetical, edited by Robert Cawdry in 1604; An English Expositor,edited by John Bullokar in 1616; An English Dictionary,edited by Elisha Coles in 1776; A New English Dictionary, edited by John Kersey in 1702; An Universal Etymological English Dictionary, edited by Nathan Bailey in 1721; Dictionarium Britannicum, also edited by Nathan Bailey in 1730; Dictionary,edited by Samuel Johnson in 1755; A Dictionary of the English Language, edited by Joseph Worcester in 1860; New Standard Dictionary of the English Language, edited by Funk and Wagnalls in 1913; and Webster's Third New International Dictionary in 1961 and Oxford English Dictionary in 1989.The Chinese dictionaries including: Erya, about 200 B.C.;Shuowen Jiezi, edited by XuChen in 100; Shenlei, in 239; Yupian,in 543; Tangyun, in 751;Leipian, in 1066; Zihui, in 1615; Kangxi Zidian in 1716; Zhonghua Dazidian in 1915; Zhongwen Dazidian in 1968; and Hanyu Da Zidian, in 1986.