Commun. Comput. Phys.
doi: 10.4208/cicp.190909.161209a

Vol. **8**, No. 3, pp. 690-700
September 2010

Comparisons of the English and Chinese Language Networks: Many Similarities and Few Differences

Dujuan Wang[1,*], Ru Wang[2] and Xu Cai[1]

[1] *Complexity Science Center, Institute of Particle Physics, Central China Normal University, Wuhan 430079, China.*
[2] *College of Information Science and Engineering, Huaqiao University, Quanzhou 362021, China.*

Received 19 September 2009; Accepted (in revised version) 16 December 2009

Communicated by Dietrich Stauffer

Available online 20 April 2010

Abstract. With words as nodes, and a link exits between two neighboring words, the weighted directed English and Chinese written human language networks are constructed from one English novel and two Chinese ones. We hereby analyze in detail the topological structure of them, in order to clarify their similar and different statistical properties. The empirical results show that the English and Chinese language networks all possess the shifted power law (SPL) degree distribution, the small-world property and the hierarchical structure, the connections among the words have positive assortativity coefficient and reciprocal characteristics. We also investigate the features of the strength and the centrality, which describe the importance of a specific word. Furthermore, considering the growth properties of the language networks and part of topological property, we find that the English written human language network grows slower than the Chinese one, which implies different mechanisms of the English and Chinese languages.

AMS subject classifications: 05C82, 91D30, 68M10

PACS: 89.75.-k, 89.75.Fb, 43.71.Sy

Key words: Topological structure, shifted power law, correlation, growth property.

1 Introduction

The past few years have witnessed people's great interest with regard to the complex networks. By investigating many real world networks, the small-world behavior [1, 2] and the scale-free property [3] were successfully confirmed, typical examples comprise

*Corresponding author. *Email addresses:* wangdj@iopp.ccnu.edu.cn (D. Wang), wr0124@gmail.com (R. Wang), xcai@mail.ccnu.edu.cn (X. Cai)

the World Wide Web [4], the collaboration network [5], the public transportation networks [6] and the graph of human language [7, 8], etc. As well known, the characterization of the topological structure [9–11] of a network is the basic factor to analyze its intrinsic functions and dynamics [12–16]. These empirical analyses have inspired people to probe the universality of the real world systems, and thus to provide an appropriate framework [17, 18] for developing techniques and models of the complex networks.

Composed of a number of words, novels and poems are simply normal examples of written human language networks in nature, and thus can be studied from the aspect of the complex network theory. Caldeira and Lobao [19] study the structure of meaningful concepts in written texts, they find the small-world effect as well as the scale-free structures. Li and Zhou [20] emphasize the Chinese character structure, supposing that the radical is the vertex and two vertices are linked if they can form a character or a part of it. Their work shows that the character networks also display the small-word property and the non-Poisson distribution. Masucci and Rodgers [21] investigate the English novel named 1984, they find the existence of different functional classes of vertices, the significance of the second order vertex correlations in the network architecture.

The previous works are of great importance to understand the nature of the written human language networks. However, to our best knowledge, they just concentrate on one language, and there are no comparisons about the characteristics of different languages. Therefore, in this paper, our main purpose is to investigate the similarities and differences between English and Chinese written human language networks. We select three novels [22] as our empirical objects, which include a Chinese one named *"A Q Zheng Zhuan"* (AQC) written by Lu Xun in 1921, the English version *"The true story of Ah'Q"* (AQE) translated by Yang Hsien-yi in 1960, and another Chinese one entitled *"Kun Lun Shang"* (KLS) written by Bi Shumin in 1986.

In our studies, word represents the vertex, an edge exists between two vertices if they are neighbors, and the edge directs from the former word to the latter one. Neglecting the punctuation marks and the paragraph gaps, we construct the weighted directed English and Chinese written human language networks, the system sizes are 21118, 17204, 23270, the number of different words are 1553, 2661, 2048 for these three networks respectively.

The whole text is organized as follows: we show the topological property of the networks such as degree distribution and clustering in Section 2. Section 3 presents the weighted network. Section 4 depicts the centrality and betweenness measures. In Section 5, we exhibits the growth properties of the networks. Conclusion and discussion are given in Section 6.

2 The topology of network

The foremost quantity that describes the characteristic of the network is the degree distribution. In order to reduce the statistical errors arising from the limited system size, we introduce the Pareto distribution, which is regarded as the same thing as Zipf, power-