Physica A 420 (2015) 315–323

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Co-occurrence network analysis of Chinese and English poems

Wei Liang [a,*], Yanli Wang [b], Yuming Shi [b], Guanrong Chen [c]

[a] School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan 454000, China
[b] Department of Mathematics, Shandong University Jinan, Shandong 250100, China
[c] Department of Electronic Engineering, City University of Hong Kong, Hong Kong Special Administrative Region

HIGHLIGHTS

- There are 34 Chinese networks with clustering coefficients being zero.
- More Chinese networks have scale-free property and hierarchical structure.
- 12 Chinese networks are assortative, but all the English networks are disassortative.
- The "M" shape appears in each of the spectral densities of the English networks.
- The triangle-like shape arises in the spectral densities of the Chinese networks.

ARTICLE INFO

Article history:
Received 31 May 2014
Received in revised form 29 October 2014
Available online 13 November 2014

keywords:
Language
Poem
Co-occurrence network
Assortativeness
Spectral analysis

ABSTRACT

A total of 572 co-occurrence networks of Chinese characters and words as well as English words are constructed from both Chinese and English poems. It is found that most of the networks have small-world features; more Chinese networks have scale-free properties and hierarchical structures as compared with the English networks; all the networks are disassortative, and the disassortativeness of the Chinese word networks is more prominent than those of the English networks; the spectral densities of the Chinese word networks and English networks are similar, but they are different from those of the ER, BA, and WS networks. For the above observed phenomena, analysis is provided with interpretation from a linguistic perspective.

1. Introduction

Geoffrey Chaucer – commonly considered as the father of English poems – was famous for using the "heroic couplet", which marked the beginning of English metrical poems. Various forms of old English metrical poems were completed and fixed in the 15–16th centuries. English free verses were developed in the 18th century. During this period, the old English metrical poems were changed under the influence of free verses and evolved into new English metrical poems. Meanwhile, English prose poems were quickly developed in the 19th century and finally evolved into today's styles of free verses, new metrical poems, and prose poems which are three pillars of the modern English poems [1,2].

After 1919, some Chinese poets combined the artistic techniques of poems existing in foreign languages, especially English, and the ancient Chinese poems, to finally create the modern Chinese poems, which have three main styles: metrical

* Corresponding author.
E-mail address: wliang@hpu.edu.cn (W. Liang).

http://dx.doi.org/10.1016/j.physa.2014.10.092
0378-4371/

poems, free verses, and prose poems [3]. Although poems of foreign languages are important resources for the formulation of the modern Chinese poems, their linguistic properties have many differences. Recently, the similarities and differences among Chinese poems and other literary styles have been studied, with some interesting phenomena observed, in Ref. [4].

Researchers, from a pure literary perspective, have comprehensive understanding in the similarities and differences between modern Chinese and English poems today [5–11]. Since poems have linguistic structures thereby forming various networks of words (and characters), it is also very natural to study poems from a networking perspective. Can the similarities and differences between modern Chinese and English poems be reflected in their corresponding word networks? Can insightful conclusions be made by studying such networks? What are the similarities and differences between English poems and other literary styles of poems from a network perspective? For the aforementioned four types of English poems, can their different writing characteristics be reflected in the statistical parameters of their word networks? The present paper attempts to address such interesting questions.

In this paper, 100 English articles of old metrical poems, new metrical poems, free verses, and prose poems are collected, respectively, from which 404 English word co-occurrence networks are constructed. Furthermore, to confirm the results concluded from analyzing the above 404 networks, 8 articles having both Chinese and English versions are selected for each type of Chinese and English poems, where the English (Chinese) versions are translated from their corresponding Chinese (English) versions. For these poems, their corresponding co-occurrence networks, a total of 168 networks, are constructed. Their statistical parameters are computed and analyzed. It is found that most of the networks have small-world feature; Chinese networks are more likely to have scale-free property; more Chinese networks but less English networks present hierarchical organizations; all the networks are disassortative, and the disassortativeness of the Chinese word networks is more prominent than the English networks; the spectral densities of the Chinese word networks and English networks are similar, yet they are different from those of the ER (Erdös–Rényi), BA (Barabási–Albert), and WS (Watts–Strogatz) networks.

2. Constructing co-occurrence networks

For English poems: old metrical poems (15th Century–19th Century), new metrical poems (19th Century–now), free verses (19th Century–now), and prose poems (19th Century–now), denoted by EOP, ENP, EFV, and EPP, respectively, 100 articles in each type are randomly selected from Ref. [12]. In addition, to probe further into the results concluded from analyzing the networks constructed using the above articles, 8 articles having both Chinese and English versions are collected for each type of Chinese poems (metrical poems, free verses, and prose poems, denoted by CMP, CFV, and CPP, respectively) and English poems, where the English (Chinese) versions are translated from their corresponding Chinese (English) versions.

A character co-occurrence network and a word co-occurrence network are constructed from each single Chinese article, and a word co-occurrence network is constructed from each single English article, denoted as C–C-network, C–W-network, and E–W-network, respectively. In a character (word) co-occurrence network based on a given article, nodes denote characters (words); two characters (words) are linked by an edge if they are adjacent to each other in a sentence of a poem.

In this paper, networks are constructed from not only single poems but also incorporated poems (combining together 100 poems of the same type). For convenience, networks constructed from single and incorporated English (Chinese) poems are denoted as S-EPNs and I-EPNs (S-CPNs and I-CPNs), respectively.

3. Main results

In this section, we explore the similarities and differences between Chinese and English poems, English poems and other literary styles, and also the aforementioned four types of English poems.

A total of 400 S-EPNs are constructed from single English poems. Furthermore, for four incorporated articles, 4 I-EPNs are constructed. Key parameters of these networks are computed, and the average values of the S-EPNs and the values of the I-EPNs are obtained, as summarized in Table 1. In addition, in order to compare English poems with Chinese poems and other literary styles, which are chosen from Refs. [4,13], basic parameters of these networks are obtained as listed in Tables 1 and 2.

3.1. Small-world feature

Recall [14] that the average shortest path length of a network with N nodes is defined as $L = 2\sum_{i>j} d_{ij}/N(N-1)$, where d_{ij} means the shortest path length that connects nodes i, j. The clustering coefficient of node i is defined as $C_i = 2E_i/k_i(k_i-1)$, where k_i is the degree of node i, and E_i is the number of the existing edges among the neighbors of i. The clustering coefficient of the whole network is the average of C_i ($1 \le i \le N$), denoted by C. A network is said to have a *small-world* feature if $L \approx L_r$ and $C \gg C_r$, where L_r is the average shortest path length and C_r is the clustering coefficient of the ER random graph [15] with the same numbers of nodes and edges.

For the 400 S-EPNs, most of them have $L \approx L_r$ and $C \gg C_r$; that is, they exhibit the small-world feature. Nevertheless, there are 27 networks with $0 < C < C_r$. These networks appear in the EOP, ENP, and EFV, where their numbers are 19, 7, and 1, respectively. However, there are 34 networks with $C = 0$ in the S-CPNs [4], while this case does not appear in the S-EPNs.