Contents lists available at ScienceDirect

Physica A

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

Comparison of co-occurrence networks of the Chinese and English languages

Wei Liang[a], Yuming Shi[a,*], Chi K. Tse[b], Jing Liu[c], Yanli Wang[a], Xunqiang Cui[a]

[a] *Department of Mathematics, Shandong University, Jinan, Shandong 250100, China*
[b] *Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China*
[c] *State Key Laboratory of Software Engineering, Wuhan University, Wuhan, Hubei 430072, China*

ARTICLE INFO

Article history:
Received 26 May 2009
Available online 4 August 2009

PACS:
89.75.Da
89.75.Fb

Keywords:
Language
Co-occurrence network
Small-world network
Scale-free network

ABSTRACT

Co-occurrence networks of Chinese characters and words, and of English words, are constructed from collections of Chinese and English articles, respectively. Four types of collections are considered, namely, essays, novels, popular science articles, and news reports. Statistical parameters of the networks are studied, including diameter, average degree, degree distribution, clustering coefficient, average shortest path length, as well as the number of connected subnetworks. It is found that the character and word networks of each type of article in the Chinese language, and the word network of each type of article in the English language all exhibit scale-free and small-world features. The statistical parameters of these co-occurrence networks are compared within the same language and across the two languages. This study reveals some commonalities and differences between Chinese and English languages, and among the four types of articles in each language from a complex network perspective. In particular, it is shown that expressions in English are briefer than those in Chinese in a certain sense.

1. Introduction

Complex networks have attracted a great deal of interest since the publication of the works of Watts and Strogatz [1] and Barabási and Albert [2]. Recently, complex network theory has been widely used to study some behaviors of complex systems in the real world such as World Wide Web and Internet [3–6], biological networks [7], collaboration networks [8], and public transport networks [9,10]. The use of the complex network approach has been found fruitful in the analysis of a variety of complex systems.

Human languages can be studied in terms of complex network models. Recently, language networks have been constructed with different criteria for connecting words or characters, such as co-occurrence [11–13], syntactic dependency [14], and semantic dependency [15–17]. These networks exhibit the small-world or scale-free feature, or both. The study of language networks has also been applied in language learning and its evolution [18,19], quantification of language characteristics (e.g. Zipf's law [20]), and comparative study among two or more languages [21,22].

There are at least 6800 different languages in the world [23]. The Chinese and English languages are two of the mostly spoken ones. For these two languages, word co-occurrence networks have been widely studied [11–13] and have been shown to exhibit small-world and scale-free features. Sentences in Chinese are formed by characters and words, while sentences in English are formed by words. Character co-occurrence networks can be constructed in a likewise manner as in the construction of word co-occurrence networks. As yet, no study has been performed on the character networks

[*] Corresponding author.
E-mail addresses: liangwei5321@yahoo.com.cn (W. Liang), ymshi@sdu.edu.cn (Y. Shi), encktse@polyu.edu.hk (C.K. Tse), j_liu@whu.edu.cn (J. Liu).

0378-4371/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2009.07.047

except for our conference paper [24]. In the existing literature, study has focused on a single network that is constructed from a large number of articles, which are selected from tagged corpus, wordnet, online English dictionary, etc. However, a character co-occurrence network and a word co-occurrence network can be constructed from a single Chinese article, and a word co-occurrence network can be constructed from a single English article. Do these networks still exhibit small-world and scale-free features? Can useful conclusions be made by comparing network parameters corresponding to two or more languages from a network perspective? In order to answer these questions, we have constructed 114 networks from collections of 53 Chinese articles, including essays, novels, popular science articles, news reports, and 4 concatenated articles of each type [24]. We found that these character and word co-occurrence networks are qualitatively equivalent, i.e., they exhibit small-world and scale-free features.

In this paper, based on our previous work on the Chinese language [24], we further study the commonalities and differences between Chinese and English, and among four types of articles, i.e., essays, novels, popular science articles, and news reports, in each language from a complex network perspective. In order to achieve this goal, 200 Chinese articles and 200 English articles are selected. A character co-occurrence network and a word co-occurrence network are constructed from a single Chinese article, and a word co-occurrence network is constructed from a single English article. Therefore, 400 Chinese character and word co-occurrence networks and 200 English word co-occurrence networks are constructed. Furthermore, in order to confirm the results concluded from analyzing the above 600 networks, 10 articles including 5 essays and 5 novels which have both Chinese and English versions are selected, where the Chinese versions are translated from their corresponding English versions, and their corresponding networks are constructed. All these networks are treated as undirected and unweighted graphs. Their statistical parameters are studied, including diameter, average degree, degree distribution, clustering coefficient, average shortest path length as well as the number of connected subnetworks. They are shown to exhibit scale-free and small-world features. We compare the statistical parameters of these networks in the same language and across the two languages and find that some statistical parameters of co-occurrence networks of different languages and different article types are almost identical while others can be quite different. This study reveals some commonalities and differences between the Chinese and English languages, and among different types of articles in each language from a complex network perspective. In particular, our empirical results show that expressions in English are briefer than those in Chinese in a certain sense.

2. Some basic concepts of complex networks

A *network* G is a set of nodes V with edges E, denoted by $G = (V, E)$. Suppose that a network with N nodes is undirected and unweighted. The *degree* of node i is the number of edges that the node has, denoted by k_i. The average degree of the network is defined by $\langle k \rangle = \sum_{i=1}^{N} k_i/N$. A network is said to be *connected* if for any two nodes in the network there is at least a path to connect these two nodes. Given two nodes $i, j \in V$, let d_{ij} be the shortest path length that connects them. Suppose that the network is connected, the diameter of the network is defined by

$$D = \max_{1 \le i,j \le N} d_{ij},$$

and the *average shortest path length* of the network is defined as

$$L = \frac{2 \sum\limits_{i>j} d_{ij}}{N(N-1)}.$$

The *clustering coefficient* C_i of node i is the probability that any two neighbors of node i are also connected to each other, i.e.,

$$C_i = \frac{2E_i}{k_i(k_i-1)},$$

where E_i is the number of the actual edges among the neighbors of node i. The clustering coefficient of the whole network is the average of C_i $(1 \le i \le N)$, denoted by C.

A random graph can be obtained by linking pairs of nodes with some probability. One of the most important random graphs is the Erdös–Rényi graph, which has a binomial degree distribution that can be approximated by a Poisson distribution [25]. For an Erdös–Rényi graph with an average degree $\langle k \rangle$, its average shortest path length is $L_r \approx lnN/ln\langle k \rangle$ and its clustering coefficient is $C_r \approx \langle k \rangle/(N-1)$. A network is said to be a *small-world network* if its average shortest path length $L \approx L_r$ and its clustering coefficient $C \gg C_r$.

Degree distribution $p(k)$ is one of the most important statistical characteristics of a network, which is defined as the probability that a randomly chosen node in the network has exactly degree k. If $p(k)$ satisfies the power-law degree distribution:

$$p(k) \propto k^{-\gamma},$$

where γ is a positive constant, then the network is said to be *scale free.*

Remarks: For convenience, we set $d_{ij} = 0$ if there is no connection between nodes i and j of a network, and $C_i = 0$ if there is no connection between node i and other nodes of a network. Then D, L, and C for a non-connected network can be likewise defined.

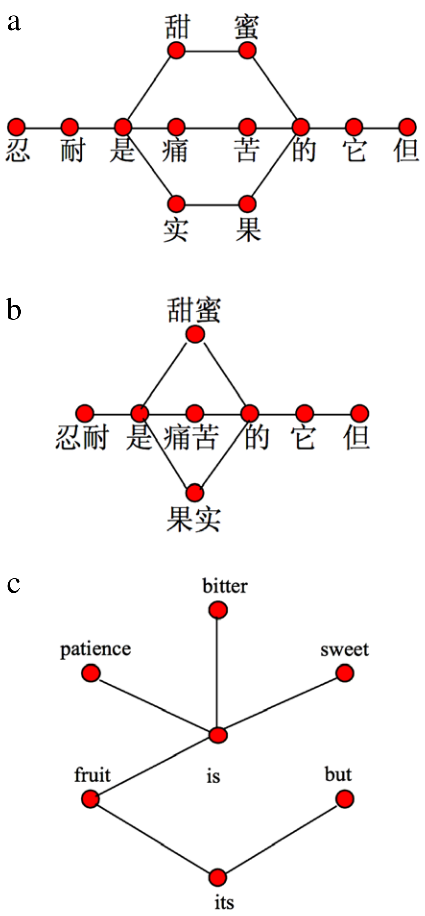

Fig. 1. Co-occurrence networks of the sample sentences with nodes being (a) Chinese characters; (b) Chinese words; (c) English words.

3. Construction of networks of Chinese and English languages

In English, most words have a meaning, whereas in Chinese, a group of characters forms a Chinese word that has a semantic meaning in general. In order to study the commonalities and differences between the Chinese and English languages, and among the four types of articles in a same language (essays, novels, popular science articles, and news reports, denoted by E, N, P, and R, respectively), 200 Chinese articles and 200 English articles are selected. A character co-occurrence network and a word co-occurrence network are constructed from a single Chinese article, and a word co-occurrence network is constructed from a single English article, denoted by C-C-network, C-W-network, and E-W-network, respectively. Therefore, 200 C-C-networks, 200 C-W-networks, and 200 E-W-networks are constructed. In addition, in order to probe further into the results concluded from analyzing the above 600 networks, 10 articles including 5 essays and 5 novels which have both Chinese and English versions are selected, where the Chinese versions are translated from their corresponding English versions, and their corresponding networks (C-C-networks, C-W-networks, and E-W-networks) are constructed. 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 occur consecutively within at least one sentence. For simplicity, these networks are treated as undirected and unweighted graphs. For example, consider the following Chinese sentence:

忍耐是痛苦的，但它的果实是甜蜜的。

Its English expression is:

Patience is bitter, but its fruit is sweet.

The corresponding C-C-network, C-W-network, and E-W-network are shown in Fig. 1.

In each language, we have collected 50 samples of each type of article, with varying lengths. Here, the length of an article is the number of characters (words), including repeated characters (words). The length of an article may have some impact on its statistical parameters. In order to provide a fair comparison, the average lengths of articles of the same type in Chinese and English languages are almost equal.

Table 1
Average values of statistical parameters of co-occurrence networks for different types of articles in Chinese and English. In the first column, E, N, P, and R denote essays, novels, popular science articles, and news reports, respectively. In the second column, the first letter C stands for Chinese and E for English; the second letter C stands for character and W for word; the third letter W represents the whole network and C the largest connected subnetwork.

Style	Network	Length	N	E	D	$\langle k\rangle$	L/L_r	$C\%/C_r\%$	γ
E	C-C-W	1311	426	821	11	3.80	3.82/4.55	6.47/0.94	2.35
	C-C-C		422	819	11	3.83	3.82/4.52	6.53/0.96	2.35
	C-W-W		415	611	12	2.93	4.13/5.62	4.67/0.77	2.86
	C-W-C		396	603	12	3.02	4.13/5.41	4.88/0.84	2.83
	E-W-W	1142	440	826	10	3.62	3.61/4.70	10.61/0.97	2.74
	E-W-C		431	824	10	3.69	3.61/4.62	10.82/1.01	2.74
N	C-C-W	5226	799	2584	9	5.97	3.40/3.83	11.17/0.82	1.98
	C-C-C		795	2583	9	6.01	3.40/3.81	11.23/0.83	1.98
	C-W-W		1006	2151	11	4.02	3.69/5.01	9.15/0.50	2.49
	C-W-C		977	2140	11	4.11	3.69/4.88	9.41/0.54	2.48
	E-W-W	5224	1314	3199	9	4.69	3.29/4.60	17.62/0.45	2.55
	E-W-C		1292	3195	9	4.77	3.29/4.54	17.92/0.47	2.54
P	C-C-W	995	313	589	11	3.72	4.05/4.42	5.20/1.24	2.17
	C-C-C		309	587	11	3.75	4.05/4.38	5.20/1.27	2.17
	C-W-W		284	434	13	3.03	4.32/5.25	4.75/1.15	2.60
	C-W-C		268	427	13	3.13	4.32/4.99	4.96/1.25	2.58
	E-W-W	961	426	734	11	3.32	3.92/5.04	8.15/0.87	2.77
	E-W-C		415	731	11	3.40	3.92/4.90	8.38/0.92	2.76
R	C-C-W	788	283	475	13	3.30	4.67/4.76	3.09/1.20	2.28
	C-C-C		280	473	13	3.33	4.67/4.71	3.13/1.24	2.28
	C-W-W		232	320	15	2.67	4.94/5.56	3.33/1.23	2.74
	C-W-C		218	305	15	2.77	4.95/5.27	3.56/1.38	2.71
	E-W-W	748	343	581	10	3.35	3.91/4.83	7.16/1.02	2.70
	E-W-C		337	579	10	3.40	3.91/4.75	7.30/1.05	2.70

For Chinese articles, the lengths fall in the ranges 613–2170 for essays (E), 797–10 387 for novels (N), 470–2512 for popular science articles (P), and 460–1613 for news reports (R), and their average lengths are 1311 (E), 5226 (N), 995 (P), and 788 (R). For English articles, the lengths fall in the ranges 454–4542 (E), 610–13 260 (N), 459–2980 (P), and 450–1941 (R), and their average lengths are 1142 (E), 5224 (N), 961 (P), and 748 (R).

In addition, for the 5 essays and 5 novels having both Chinese and English versions, the lengths of the Chinese articles fall in the ranges 1036–3322 (E) and 2033–13 143 (N), and their average lengths are 1770 (E) and 5222 (N); and the lengths of the English articles fall in the ranges 628–1874 (E) and 1244–9180 (N), and their average lengths are 1141 (E) and 3330 (N).

4. Empirical results with analysis

In this section, we report our main results on commonalities and differences between the networks constructed from Chinese and English articles, and among the four types of articles in each language.

4.1. *Comparison between Chinese and English languages*

In this subsection, we explore the commonalities and differences between Chinese and English articles. Specifically, we compute some network parameters for all the networks constructed, including diameter, average degree, degree distribution, clustering coefficient, average shortest path length, and the number of connected subnetworks. The average parameter values for the different types of articles are reported in Table 1.

4.1.1. *Scale-free degree distribution*

We first analyze the degree distributions in the 600 co-occurrence networks. For these networks, 100% of C-C-networks, 99% of C-W-networks, and 99.5% of E-W-networks, i.e., 597 networks, exhibit power-law degree distributions.

Fig. 2 shows the degree distributions plotted in the log–log scale for typical C-C-networks, C-W-networks, and E-W-networks for the four different types of articles. Fig. 3 shows the degree distributions of the three networks whose degree distributions are not power law. In the 597 networks that display power-law degree distributions, the power-law exponents, γ, fall in the ranges 1.45–2.98 for the C-C-networks, 1.73–3.25 for the C-W-networks, and 2.10–3.47 for the E-W-networks. The values of γ have been computed by least-square-error estimation. In these networks, the average values of γ in the C-W-networks are larger than those in the C-C-networks in general and are closer to those in the E-W-networks (see Table 1).

It is widely known that γ is about 3 in the BA scale-free network [2], where the scale-free feature is a result of growth and preferential attachment. However, in the 597 networks constructed, 93.65% of power-law exponents are less than 3 and the rest are larger than 3. On the other hand, all average values of γ of these 597 networks are less than 3 (see Table 1). Therefore,

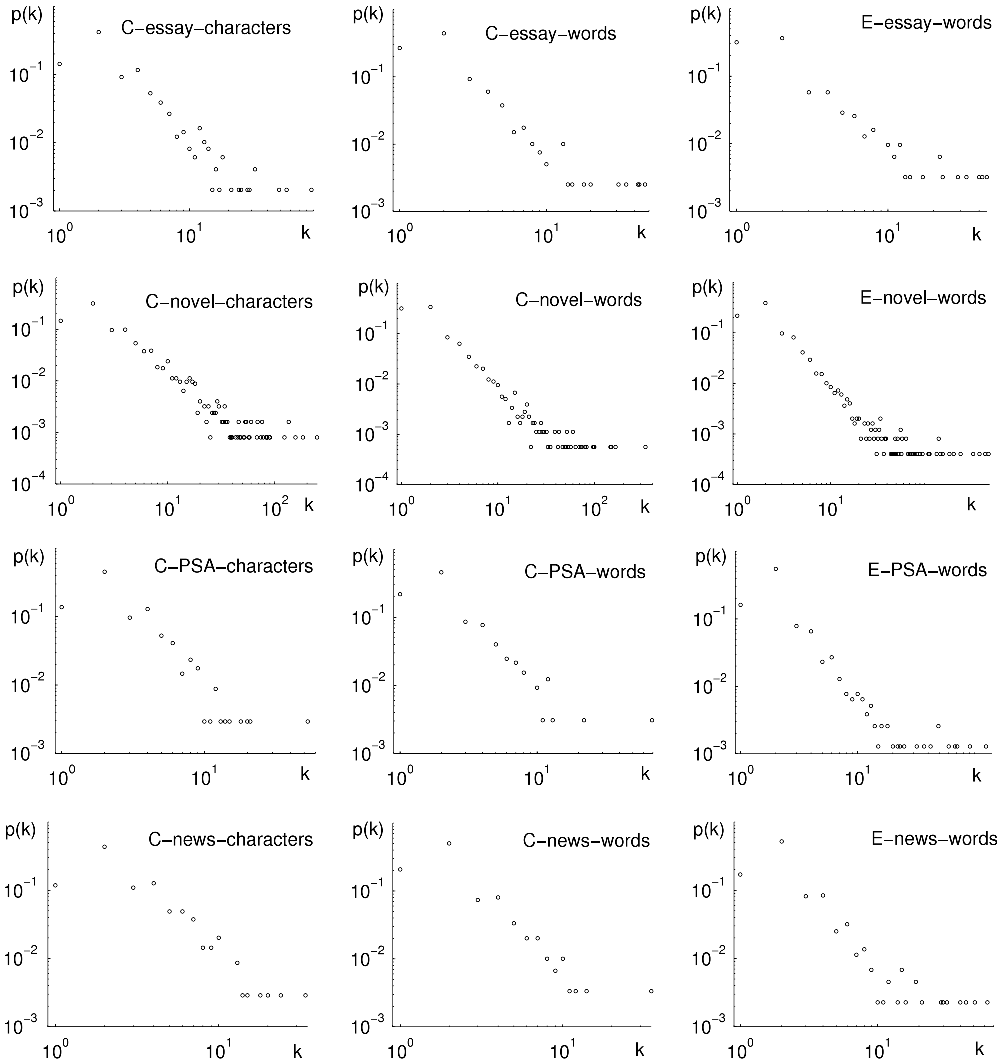

Fig. 2. Degree distributions of co-occurrence networks for essays, novels, popular science articles, and news reports in the log–log scale, where C represents Chinese and E represents English.

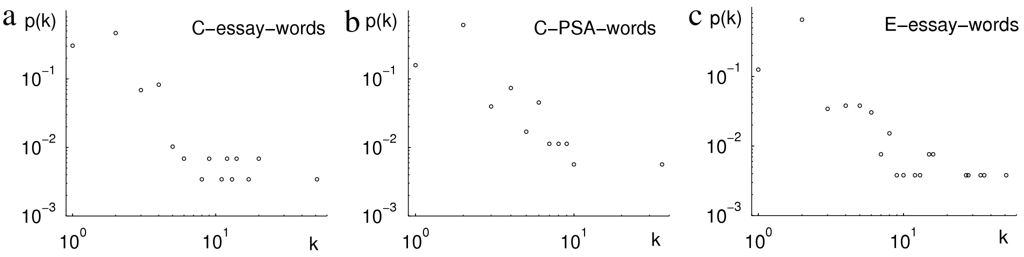

Fig. 3. Degree distributions of the three networks in the log–log scale, which are not like power law.

there may be other link attachment processes contributing to the scale-free feature, apart from preferential attachment. In these networks, the degree distribution of every network can be fitted by a straight line in the log–log scale. This is different from the case of the word co-occurrence networks reported in Refs. [11,12], where the degree distribution of the network is fitted by two lines in the log–log scale, i.e., piecewise scale-free distribution with two power-law exponents.

A power-law degree distribution of a network implies that there are relatively few number of nodes (hubs) having very large number of connections, while most nodes have few connections in the network [2]. According to the statistics observed in the networks constructed, the nodes with large number of connections are “a”, “the”, and “of” in the English language networks, and “的”, “了”, and “是” in the Chinese language networks. Also, some functional words and words that are closely related to the topics of the articles are highly connected nodes.

The degree distributions of most networks follow a power law. There are only three networks whose degree distributions do not follow a power law (see Fig. 3). One of the three special networks belongs to the 50 C-W-networks from essays, one to the 50 C-W-networks from popular science articles, and one to the 50 E-W-networks from essays. The numbers of nodes in these three networks are 292, 177, and 263. If the 50 networks of each type of article are ranked according to the number of nodes in the network, then these three networks are ranked 6th, 3rd, and 8th among the 50 networks of the same type. Meanwhile, the lengths of articles are, correspondingly, 782, 537, and 621. If the 50 articles of each type are ranked according to the length of the article, then the corresponding three articles are ranked 6th, 2nd, and 8th among the 50 articles of the same type. We also observe that if a network has more nodes, its degree distribution resembles more closely a power-law distribution. Therefore, the three networks that do not display power-law degree distributions are networks with relatively fewer number of nodes, and hence may exhibit irregularities. We may therefore conclude that co-occurrence networks have power-law degree distributions. This phenomenon reflects that both Chinese and English languages are self-organizing systems like many real-world networks.

4.1.2. Connectivity of networks

As discussed in Section 4.1 in Ref. [24], we also found that 91.33% of the networks are disconnected, and the statistical parameters of the whole network are nearly the same as those of its largest connected subnetwork (see Table 1).

In the 600 networks constructed, the average number of connected subnetworks for the four different types of articles are 3, 4, 3, 3 (E, N, P, R) for the C-C-networks; 11, 24, 10, 8 for the C-W-networks; and 8, 21, 9, 5 for the E-W-networks. Thus, we observe that the C-C-networks are better connected than the C-W-networks (see Fig. 4), and the E-W-networks are better connected than the C-W-networks, but are less connected than the C-C-networks.

4.1.3. Small-world effect

We now analyze the small-world property in our networks. As discussed in Section 2, small-world property applies only to connected networks. Therefore, we only consider the average shortest path lengths and clustering coefficients of the largest connected subnetworks although they are nearly the same as the respective whole networks, as discussed in Section 4.1.2.

Based on our empirical results shown in Table 1, we have observed that all the 600 networks have small diameters D and small average shortest path lengths L. The diameters of all these networks are below 15. The ranges of L of the C-C-networks, C-W-networks, and E-W-networks are 2.76–5.84, 2.88–6.31, and 3.04–4.56, respectively. As shown in Table 1, the average values of L for the four types of articles are all below 5. This means that for any two nodes in any given network, we need fewer than five nodes on average to connect them. The reason why the values of L are so small may be related to the existence of hubs, which play a bridging role in connecting two different nodes of the networks. Therefore, in the cognition process, although a huge number of characters (words) are stored in the human brain, each character (word) can be reached with few intermediate characters (words) on average. This assures the high processing speed during speech production. In addition, the C-C-networks have smaller L than the C-W-networks, and the E-W-networks have smaller L than the two kinds of Chinese language networks in each type of article.

Let L_r be the average shortest path length of the corresponding random network. The ranges of L_r for the C-C-networks, C-W-networks, and E-W-networks are 2.72–5.60, 3.28–6.69, and 3.81–6.03, respectively. The average values of L_r for the four types of articles are listed in Table 1. It is shown that L is very close to L_r, i.e., $L \approx L_r$ for all the 600 networks constructed.

All 600 networks have small clustering coefficients C. The ranges of C for the C-C-networks, C-W-networks, and E-W-networks are 0.0143–0.262, 0.0013–0.242, and 0.029–0.264, respectively. The ranges of the clustering coefficients of the corresponding random networks C_r for the C-C-networks, C-W-networks, and E-W-networks are 0.005–0.0196, 0.0021–0.0226, and 0.0018–0.0224, respectively. The average values of C and C_r for the four types of articles are listed in Table 1. Obviously, the clustering coefficient of each network is much larger than that of the corresponding random network, i.e., $C \gg C_r$. On the other hand, the clustering coefficients C of the C-C-networks and C-W-networks are almost the same, and the E-W-networks have larger values of C than the two kinds of Chinese language networks for each type of article.

Based on the above analysis, we can conclude that each network constructed here has a small average shortest path length and is highly clustered. Therefore, each network constructed here exhibits a small-world feature. Furthermore, the average values of L of the English language networks are smaller than those of the two kinds of Chinese language networks. Meanwhile, the average values of C of the English language networks are larger than those of the two kinds of Chinese language networks. This implies that less words are needed to connect any two words in the English language than in the Chinese language. In this sense, expressions in English are briefer than in Chinese.

4.1.4. Co-occurrence networks from articles in both Chinese and English languages

In order to find out if the results obtained in Sections 4.1.1–4.1.3 are universal, we select 10 articles including 5 essays and 5 novels which have both Chinese and English versions, and study the three kinds of networks, namely, C-C-network, C-W-network, E-W-network. Their statistical parameters are computed and their average values are listed in Table 2.

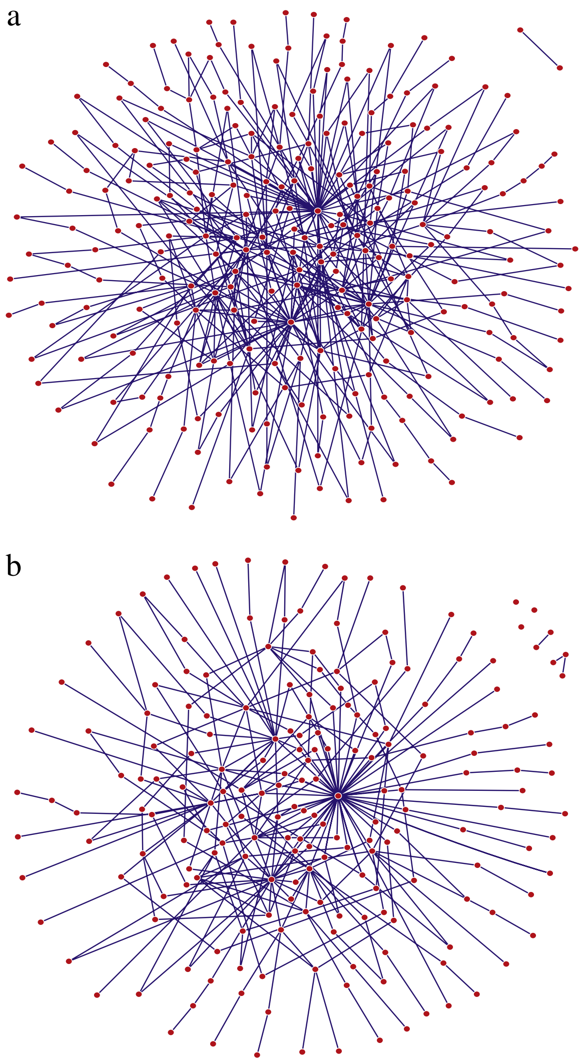

Fig. 4. Networks from an essay. (a) Chinese character network with 253 nodes and 408 edges and (b) its corresponding Chinese word network with 198 nodes and 279 edges.

Table 2
Average values of statistical parameters of co-occurrence networks for essays and novels having both Chinese and English versions, where the Chinese versions are translated from their corresponding English versions.

Style	Network	Length	N	E	D	$\langle k \rangle$	L/L_r	$C\%/C_r\%$	γ
E	C-C-W	1770	454	1060	10	4.52	3.66/4.10	8.00/1.03	2.10
	C-C-C		452	1059	10	4.53	3.66/4.09	8.03/1.05	2.10
	C-W-W		447	777	11	3.34	3.96/5.09	6.66/0.80	2.30
	C-W-C		433	772	11	3.42	3.96/4.95	6.88/0.86	2.30
	E-W-W	1141	431	829	10	3.79	3.58/4.55	11.03/0.97	2.51
	E-W-C		426	828	10	3.83	3.58/4.50	11.15/1.00	2.51
N	C-C-W	5222	813	2537	9	5.95	3.37/3.83	10.28/0.85	1.88
	C-C-C		810	2536	9	5.97	3.37/3.82	10.32/0.86	1.88
	C-W-W		956	2087	10	4.12	3.63/4.95	8.69/0.55	2.38
	C-W-C		935	2081	10	4.20	3.63/4.84	8.86/0.58	2.38
	E-W-W	3330	872	2082	9	4.39	3.35/4.63	14.34/0.58	2.45
	E-W-C		856	2079	9	4.47	3.35/4.54	14.63/0.61	2.45

Although the data in Table 2 are different from those in Table 1, all conclusions obtained in Sections 4.1.1–4.1.3 still hold. First, each network has a power-law degree distribution and exhibits a small-world feature. Second, the English language

Table 3
Values of the statistical parameters of three co-occurrence networks for the two concatenated articles, one generated from concatenating the 200 Chinese articles and the other from concatenating the 200 English articles. The data in the last line are chosen from Ref. [13].

Network	Length	N	E	D	$\langle k\rangle$	L/L_r	$C\%/C_r\%$	γ
C-C-W	415 968	4 520	96 512	8	42.70	2.49/2.24	38.07/0.95	1.19
C-W-W		23 317	119 016	10	10.21	2.95/4.33	29.02/0.04	1.88
E-W-W	403 720	25 272	143 439	10	11.35	2.85/4.17	39.97/0.04	1.89
CLN1		62 281	400 717		12.80	3.04/4.35	50.91/0.02	1.90

networks have smaller average shortest path lengths L and larger clustering coefficients C than the Chinese language networks, i.e., expressions in English are briefer than those in Chinese.

4.1.5. Three networks from concatenated Chinese and English articles

In this subsection, three networks, namely, C-C-network, C-W-network, and E-W-network, are constructed from two concatenated articles, one generated from concatenating the 200 Chinese articles and the other from concatenating the 200 English articles. Statistical parameters of the whole network are nearly the same as those of its largest connected subnetwork as discussed in Section 4.1.2. Therefore, we only compute the statistical parameters of the three whole networks in this subsection. These parameter values are shown in Table 3.

The three networks from the two concatenated articles all exhibit scale-free and small-world features. In addition, the power-law exponent γ of the C-W-network is larger than that of the C-C-network and is closer to that of the E-W-network; the clustering coefficient C of the C-C-network is larger than that of the C-W-network, and the value of C of the E-W-network is larger than those of both C-C-network and C-W-network; the average shortest path length L of the C-C-network is shorter than that of the C-W-network, and the value of L of the E-W-network is shorter than that of the C-W-network and larger than that of the C-C-network. These results are different from those obtained in Section 4.1.3 for networks constructed from single articles. To see why the value of L of the C-C-network is smaller than that of the E-W-network, we note that there are fewer nodes and more edges in the C-C-network than those of the E-W-network (see Table 3). We know that there are about 5000 Chinese characters in the Chinese language and about 5000 English words in the English language which are in common use. However, we observe that the number of nodes of the C-C-network is much fewer than that of the E-W-network even if the lengths of the two concatenated articles are nearly the same. This phenomenon can be explained as follows. In the English language, a word may have several derivatives. For example, a noun has its singular and plural forms, and a verb has its present tense, past tense, and past participle forms. These derivatives are different in general. These different derivatives will define different nodes in networks. However, there are no such derivatives for characters in the Chinese language. For example, “be” has the following different derivatives: “is”, “was”, “are”, “were”, “being”, and “been” in the English language. All these English words correspond to only one Chinese character “是 ”. Therefore, several different English words may correspond to only one Chinese character.

The values of γ and L for the three networks constructed from the two concatenated articles are smaller than those of the networks from the individual articles of the same type in general, while the values of C are larger than those of the networks from the individual articles of the same type (see Tables 1 and 3). These results are the same as those obtained in Ref. [24].

Finally, we compare our results with the results obtained in Ref. [13]. The network CLN1 in Ref. [13] was constructed in the same way as the C-W-network constructed in this subsection, including the definitions of nodes and edges, undirectedness, and unweightedness. It was also constructed from a large number of articles. It exhibits both scale-free and small-world features. In particular, though CLN1 has more nodes and edges than our C-W-network, its γ and L values are fully consistent with those of the respective C-W-network (see Table 3).

4.2. Comparison among four types of articles

In this subsection, we analyze commonalities and differences among the four types of articles in each language from a complex network perspective according to the data shown in Table 1.

In the Chinese language, the networks constructed from novels have the smallest D, L, and γ, and the largest $\langle k\rangle$ and C. Also, the networks from news reports have the largest D and L, and the smallest $\langle k\rangle$ and C. Moreover, for the networks from essays and popular science articles, no clear distinction can be made. In general, the parameters of the networks from essays and popular science articles are very similar. Therefore, from the complex network perspective, novels and news reports are significantly different; moreover, essays and popular science articles share some common features.

From the Chinese language networks, we have observed the following interesting phenomenon. The parameters L and γ of the C-W-networks are larger than those of the C-C-networks for the four different types of articles. The differences in L between C-W-networks and C-C-networks for the four types of articles are 0.31 (E), 0.28 (N), 0.27 (P), and 0.28 (R) (see Table 1). The differences in γ between C-W-networks and C-C-networks for the four types of articles are 0.51 (E), 0.51 (N), 0.43 (P), and 0.46 (R) (see Table 1). We also note that these differences are very similar in magnitude among the different types of articles. Since a word is formed by a group of characters in Chinese, this phenomenon may reveal that there are some rules in the formation of words from characters.

In the English language, the networks from novels still have the smallest D, L, and γ and the largest $\langle k \rangle$ and C, and the networks from the other three types have almost the same γ. Moreover, other parameters for the news reports and popular science articles are highly consistent, whereas those of essays are significantly different.

Based on the above analysis, the networks from novels have the smallest D, L, and γ and the largest $\langle k \rangle$ and C in the four types of articles in both Chinese and English. Therefore, novels are a special type of articles from a complex network perspective. Furthermore, essays and popular science articles in the Chinese language share some common features, whereas the news reports and popular science articles in the English language share some common features.

5. Conclusions

Co-occurrence networks of Chinese characters and words, and of English words, are constructed from collections of four different types of Chinese and English articles: essays, novels, popular science articles, and news reports. We have found that all the networks constructed exhibit scale-free and small-world features. Meanwhile, based on the empirical data, our analysis shows that expressions in English are briefer than those in Chinese in a certain sense. Furthermore, our study shows that there are some commonalities and differences among the four different types of articles. Novels are a special type. Essays and popular science articles in the Chinese language share some common features, whereas news reports and popular science articles in the English language share some common features. Furthermore, we have observed an interesting phenomenon that in the Chinese language, the differences in the average shortest path lengths and the power-law exponents among the four different types of articles are highly consistent in magnitude.

Acknowledgments

This research was supported by the NNSF of Shandong Province, China (grant Y2006A15) and Hong Kong Polytechnic University (grant G-YH59).

References

[1] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' network, Nature 393 (1998) 440–442.
[2] A.-L. Barabási, R. Albert, Emergence of scaling in random network, Science 286 (1999) 509–512.
[3] B.A. Huberman, P.L.T. Pirollo, J.E. Pitkow, R.M. Lukose, Strong regularities in the world wide web surfing, Science 280 (1998) 95–97.
[4] S. Lawrence, G.L. Giles, Searching the world wide web, Science 280 (1998) 98–100.
[5] S. Lawrence, G.L. Giles, Accessibility of information on the web, Nature 400 (1999) 107–109.
[6] R. Albert, A.-L. Barabási, Topology of evolving networks: Local events and universality, Phys. Rev. Lett. 85 (2000) 5234–5237.
[7] H. Jeong, B. Tombor, R. Albert, Z.N. Oltwai, A.-L. Barabási, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.
[8] M.E.J. Newman, The structure of scientific collaboration-networks, Proc. Natl. Acad. Sci. USA 98 (2001) 404–409.
[9] M. Kurant, P. Thiran, Layered complex networks, Phys. Rev. Lett. 96 (2006) 138701.
[10] Y.-Z. Chen, N. Li, D.-R. He, A study on some urban bus transport networks, Physica A 376 (2007) 747–754.
[11] R. Ferrer i Cancho, R.V. Solé, The small world of human language, Proc. R. Soc. Lond. B 268 (2001) 2261–2265.
[12] Z.-Y. Liu, M.-S. Sun, Chinese word co-occurrence network: Its small world effect and scale-free property, J. Chinese Information Processing 6 (2007) 52–58.
[13] S.-G. Zhou, G.-B. Hu, Z.-Z. Zhang, J.-H. Guan, An empirical study of Chinese language networks, Physica A 387 (2008) 3039–3047.
[14] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) 051915.
[15] M. Sigman, G.A. Cecchi, Global organization of the wordnet lexicon, Proc. Natl. Acad. Sci. 99 (2002) 1742–1747.
[16] M. Steyvers, J.B. Tenenbaum, The large-scale structure of semantic network: Statistical analysis and a model of semantic growth, Cognitive Science 29 (2005) 41–78.
[17] L. Tang, Y.-G. Zhang, X. Fu, Structures of semantic networks: How do we learn semantic knowledge, J. Southeast University 22 (2006) 413–417.
[18] M.A. Nowak, J.B. Plotkin, V.A.A. Jansen, The evolution of syntactic communication, Nature 404 (2000) 495–498.
[19] M.A. Nowak, D.C. Krakauer, The evolution of language, Proc. Natl. Acad. Sci. USA 96 (1999) 8028–8033.
[20] G.K. Zipf, Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley Press, Cambridge, MA, 1949.
[21] D. Abrams, S. Strogatz, Modeling the dynamics of language death, Nature 424 (2003) 900–900.
[22] T. Teşileanu, H. Meyer-Ortmanns, Competition of languages and their hamming distance, Int. J. Mod. Phys. C 17 (2006) 259–278.
[23] B.F. Grimes, Ethnologue: Languages of the World, 14th ed., Summer Institute of Linguistics, Dallas, TX, 2000.
[24] Y.-M. Shi, W. Liang, J. Liu, C.K. Tse, Structural equivalence between co-occurrences of characters and words in Chinese language, in: International Symposium on Nonlinear Theory and its Applications, 2008, pp. 94–97.
[25] P. Erdös, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17–60.