Physica A 393 (2014) 579–589

Contents lists available at ScienceDirect

Physica A

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

Comparison of directed and weighted co-occurrence networks of six languages

Yuyang Gao [a], Wei Liang [b], Yuming Shi [c,*], Qiuling Huang [d]

[a] School of Computer Science and Technology, Shandong University, Jinan, Shandong 250100, China
[b] School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan 454000, China
[c] School of Mathematics, Shandong University, Jinan, Shandong 250100, China
[d] School of Mathematics and Quantitative Economics, Shandong University of Finance and Economics, Jinan, Shandong 250014, China

HIGHLIGHTS

- The English word connections are denser and its expression is more flexible.
- Statistical data have shown that French and Spanish languages share many commonalities.
- Statistical data have shown that Chinese and English languages share many commonalities.
- Arabic and Russian word connections are sparse.
- Chinese word connections obey a more uniform distribution.

ARTICLE INFO

Article history:
Received 20 June 2013
Received in revised form 20 August 2013
Available online 9 September 2013

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

ABSTRACT

To study commonalities and differences among different languages, we select 100 reports from the documents of the United Nations, each of which was written in Arabic, Chinese, English, French, Russian and Spanish languages, separately. Based on these corpora, we construct 6 weighted and directed word co-occurrence networks. Besides all the networks exhibit scale-free and small-world features, we find several new non-trivial results, including connections among English words are denser, and the expression of English language is more flexible and powerful; the connection way among Spanish words is more stringent and this indicates that the Spanish grammar is more rigorous; values of many statistical parameters of the French and Spanish networks are very approximate and this shows that these two languages share many commonalities; Arabic and Russian words have many varieties, which result in rich types of words and a sparse connection among words; connections among Chinese words obey a more uniform distribution, and one inclines to use the least number of Chinese words to express the same complex information as those in other five languages. This shows that the expression of Chinese language is quite concise. In addition, several topics worth further investigating by the complex network approach have been observed in this study.

1. Introduction

Complex networks have attracted a great deal of interest because of variety and wide range of their applications. In particular, since Watts and Strogatz [1] introduced the small-world feature and Barabási and Albert [2] found the scale-free feature, a great progress has been made in the study of complex networks. Recently, the complex network theory has

* Corresponding author.
E-mail address: ymshi@sdu.edu.cn (Y. Shi).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.08.075

been widely applied to the study of some behaviors of complex systems in the real world such as World Wide Web and Internet [3–6], biological networks [7–9], collaboration networks [10,11] and public transport networks [12,13]. There have been fruitful results obtained by the complex network approach for the analysis of various complex systems.

A human language can be viewed as a complex adaptive system formed by a long-time evolution [14]. Some results have been obtained from the complex network perspective. Cancho and Solé [15] applied the complex network approach to the study of human languages in 2001. They built an English co-occurrence network and found that it has both small-world and scale-free properties. Since then, some scholars studied and analyzed language networks in different levels, including co-occurrence networks [15–19], syntactic networks [20,21], semantics networks [22–24] and conception networks [25]. These language networks exhibit either the small-world or scale-free feature, or both. The existing literature focused on a single network that was constructed from a large number of articles or a big corpus except the networks in Refs. [17–19], which were built from a single article.

A general law of languages is one of the most important topics in the study of languages [26]. There are at least 6800 different languages now being used in the world [27]. If one only studies one of them, then it is difficult to find any general laws of languages. In recent years, some scholars applied the complex network method to compare several different languages. In Ref. [21], constructing syntactic networks from large corpora of Czech, German and Romanian languages, the authors studied some structural features of the networks. In Ref. [28], the authors built syntactic networks to study 15 languages, including Arabic, Chinese, English, French and Spanish languages. In Ref. [29], the authors constructed two weighted networks from an English novel and a Chinese biography, where the weight equals the frequency of the corresponding word appearing in the text. In Refs. [30–32], the authors studied the classification of languages based on language complex networks. Recently, in order to study the commonalities and differences between the Chinese and English languages, we constructed character and word co-occurrence networks based on corpora built by a single article and concatenated article from collections of Chinese articles, including essays, novels, popular science articles and news reports [18]. We investigated some statistical parameters of the networks, including average degree, degree distribution, average shortest path length, clustering coefficient, etc.

It is known that connections among words in each language are directed. Therefore, it seems more natural and reasonable to construct directed networks to study languages. To our best of knowledge, most of language networks constructed in the existing literature are undirected, and there have been no language networks that are both weighted and directed. In addition, in comparing different languages, if the corpora used to construct networks describe same events, then these statistical data obtained by these corpora must more accurately characterize similarities and differences among these languages. However, there have been no such research works by this method except for Refs. [18,32], and in Ref. [18] only 10 articles written in both Chinese and English were selected.

The official languages used in the United Nations are Arabic, Chinese, English, French, Russian and Spanish. Most UN documents are issued in these six languages. These documents should be written by professional translators, and hence their writings should be very standard. In order to get corpora of different languages that describe same events, we select some articles from the UN secretary-general reports, each of which were written in these six languages. Based on these corpora, we construct six weighted and directed word co-occurrence networks.

We know that these six languages have experienced a long period of development. We briefly introduce them as follows.

Arabic language is a synthetic language. It belongs to the central Semitic branch of the Semitic family of the Semito-Hamitic language system. It is used as a common language in the Middle East and Northern Africa area, and currently an official language in 22 countries. More than 210 million people take it as their mother tongue. In addition, it is the religious language for the Muslims in the whole world. "Koran", the central religious text of Islam, was written in it.

Chinese language is an isolating language. So it is an analytic language. It belongs to the Sinitic family of the Sino-Tibetan language system. Chinese characters are logogram and have certain phonetic features. It consists of written and spoken languages. The ancient written language is called the classical Chinese, and the modern written language refers to the modern standard Chinese. The modern spoken language has many dialects, some of which are quite different. But the modern written language is quite unified. About 15% of the world's population speak Chinese as their native language. It is the largest language in the world.

English language is an analytic language. It belongs to the Western Germanic language branch of the Indo-European language system. It was widely propagated around the world by the British colonial activities. Because of assimilation of words from many different languages throughout history, modern English contains a very large vocabulary with complex and irregular spelling. 75 countries take it as an official language. It is the second largest language in the world. About 461 million people speak it, and over 1.01 billion people are learning it. So it is now the most powerful language in the world.

French language is an analytic language and has some characteristics of synthetic language. It belongs to the Western Romance branch of the Romance language family of the Indo-European system. It is estimated as having about 110 million native speakers and 190 million second language speakers in the worlds. It is one of the languages spoken by most people in the Romance languages.

Russian language is a synthetic language. It belongs to the Eastern Slavic branch of the Indo-European language system. It is primarily spoken in Russia and the other countries that were members of the former Soviet Union. About 240 million people speak it in the world. It is taken as an official language in Russia, Belarus, Kazakhstan, Kyrgyzstan, Republic of Transnistria, South Ossetia, Abkhazia and other former Soviet republics. It is the 8th most spoken language in the world by number of native speakers and the 5th by total number of speakers.

Spanish language is a synthetic language. It belongs to the Western Romance branch of the Romance language family of the Indo-European language system. There are about 350 million people speaking it as a native language, which mainly distribute in Latin American. It is the third most spoken language in the world and the second most spoken language by number of native speakers after Chinese language.

In order to compare characteristics of these six languages, we select 100 articles from the UN secretary-general reports issued on the United Nations website, each of which is written in these six languages, separately. Based on these corpora, we construct six weighted and directed word co-occurrence networks. We compute their statistical parameters, including article length; numbers of strongly connected subnetworks, nodes, edges, repetitions and degrees (in-degree, out-degree and total degree); distributions of numbers of repetitions, degrees and edge weights; average shortest path length; clustering coefficient, etc. It is shown that all these six networks exhibit both scale-free and small-world features. It is more important that we find several new non-trivial results. At the same time, we find several topics worth further studying by the complex network approach.

2. Construction of language networks

In order to study commonalities and differences among Arabic, Chinese, English, French, Russian and Spanish languages, 100 UN reports are selected, and each of them is written in these 6 languages, separately. The length of each report is about 20 pages. Then 6 weighted and directed word co-occurrence networks are constructed from 6 concatenated articles, each of which is combined together by the 100 reports in a same language. In a weighted and directed word co-occurrence network, nodes are words, two words are connected by a directed edge if they are adjacent to each other, the direction of the edge is pointed from the antecedent to the consequent, and the weight of the edge denotes the frequency of connection between the two words appearing in the text. For example, consider the following sentences: The corresponding Arabic expression is

وفقا لأحكام ميثاق الأمم المتحدة، فإن الغرض من الأمم المتحدة هو الحفاظ على السلم والأمن الدوليين.

The corresponding Chinese expression is

根据联合国宪章的规定，联合国的宗旨是维护国际和平与安全。

The corresponding English expression is

In accordance with the provisions of the Charter of the United Nations, the purpose of the United Nations is to maintain international peace and security.

The corresponding French expression is

En conformité avec les dispositions de la Charte des Nations Unies, le but de l'Organisation des Nations Unies est de maintenir la paix et la sécurité internationales.

The corresponding Russian expression is

В соответствии с положениями Устава Организации Объединенных Наций, целью Организации Объединенных Наций является поддержание международного мира и безопасности.

The corresponding Spanish expression is

De conformidad con las disposiciones de la Carta de las Naciones Unidas, el propósito de las Naciones Unidas es mantener la paz y la seguridad internacionales.

Their corresponding weighted and directed word co-occurrence networks are shown in Fig. 1.

3. Empirical results with analysis

In this section, we explore commonalities and differences among Arabic, Chinese, English, French, Russian and Spanish languages through analyzing data of statistical parameters. We compute the following parameters for all the networks constructed: article length; numbers of strongly connected subnetworks, nodes, edges, edge weights, repetitions and degrees (in-degree, out-degree and total degree); distributions of edge weights, repetitions and degrees; average shortest path length; diameter; and clustering coefficient. The data of statistical parameters of the whole networks and their largest strongly connected subnetworks are listed in Tables 1 and 2, respectively.

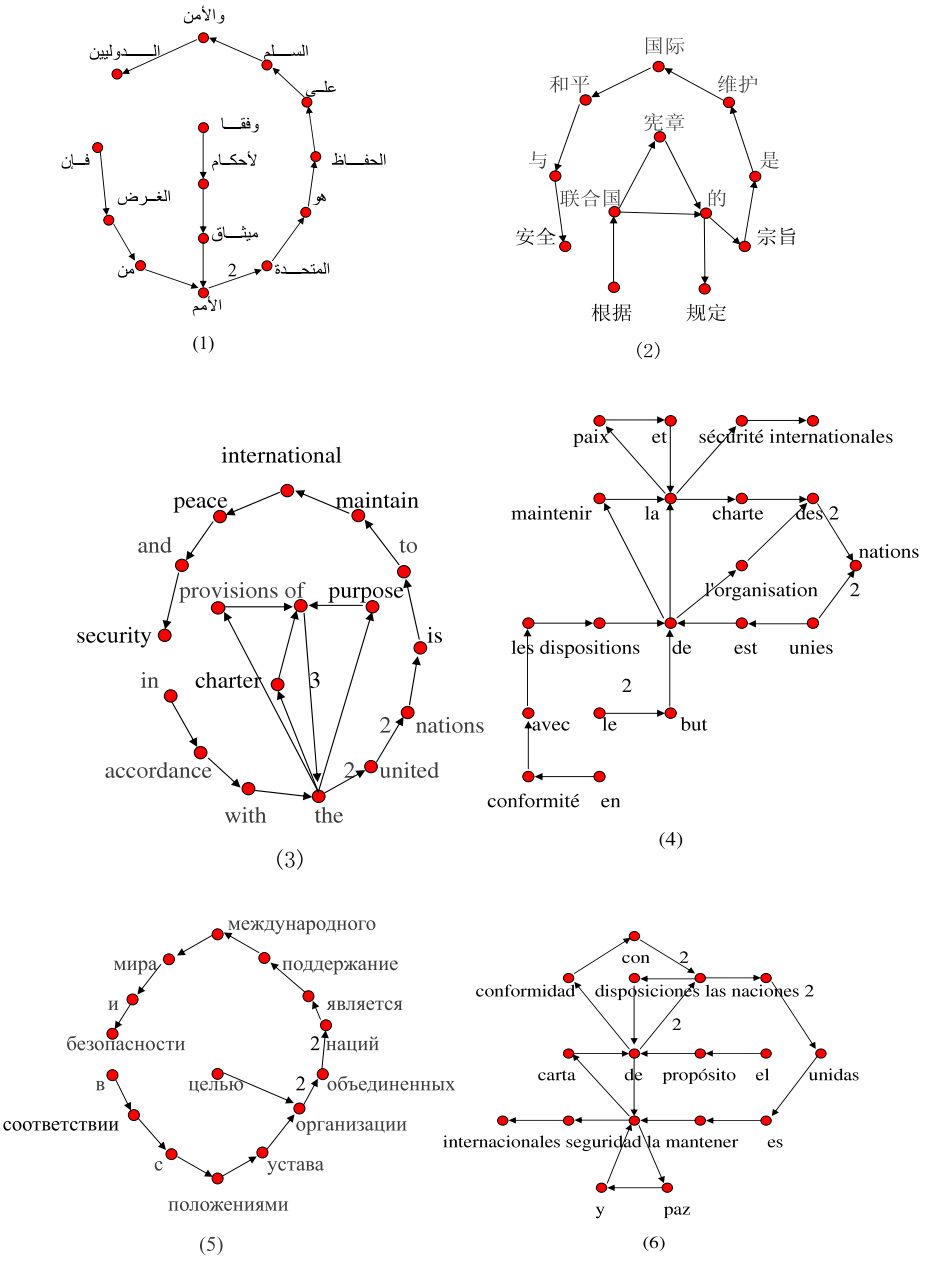

Fig. 1. Co-occurrence networks of the sample sentences: (1) Arabic, (2) Chinese, (3) English, (4) French, (5) Russian, (6) Spanish, where the numbers on the edges denote the edge weights, and they are omitted if they equal to 1.

In order to find the commonalities and differences among these six languages, we shall arrange the values of statistical parameters in order. For briefness, by A, C, E, F, R and S, the first letters of these languages, denote them, respectively, in all the inequalities and figures.

3.1. Strong connectivity of the networks

If any two nodes in a directed network are reachable to each other, then the network is said to be strongly connected. The six networks are not strongly connected. The numbers of their strongly connected subnetworks are ordered as

$$2061(E) < 2267(C) < 2877(F) < 3263(S) < 5741(A) < 6554(R). \tag{1}$$

It is shown that the English network has the best strong connectivity, followed by the Chinese network, while the strong connectivity of the Russian network is the worst, and the Arabic network is the second worst.

Table 1
Values of statistical parameters of networks for the six languages. In the first column, by SCS denote numbers of the strongly connected subnetworks.

Languages	A	C	E	F	R	S
Node	42 154	13 927	16 659	22 348	37 044	22 316
Edge	21 8557	157 453	159 946	163 010	219 503	148 379
Length	53 3268	489 841	617 948	728 315	577 896	756 856
SCS	5741	2267	2061	2877	6554	3263
Repetition	10.28	35.17	37.09	32.59	15.60	33.92
Biggest repetition	16 994	31 396	51 551	49 848	29 389	71 497
Repetition slope	1.77	1.47	1.50	1.61	1.73	1.56

Table 2
Values of statistical parameters of the largest strongly connected subnetworks for the six languages. In the first column, by LS denote the largest subnetwork.

Languages	A	C	E	F	R	S
Nodes of LS	36 412	11 661	14 599	19 472	30 490	19 053
Edges of LS	212 083	155 000	157 597	159 881	212 513	145 010
Total number of edge weight in LS	472 417	427 133	565 300	669 933	508 543	700 906
Average in-degree	5.82	13.29	10.80	8.21	6.97	7.61
Biggest in-degree	5732	3776	4381	5351	6468	5710
In-degree slope	1.92	1.58	1.81	2.09	1.89	2.00
Average out-degree	5.82	13.29	10.80	8.21	6.97	7.61
Biggest out-degree	3070	3743	4646	3573	8201	3913
Out-degree slope	1.90	1.56	1.72	1.83	1.93	1.82
Average total degree	11.65	26.58	21.59	16.42	13.94	15.22
Biggest total degree	8802	7519	9027	8924	14 669	9460
Total degree slope	1.83	1.52	1.64	1.83	1.86	1.85
Average edge weight	1.76	2.76	3.59	4.19	2.39	4.84
Biggest edge weight	2302	664	9079	9487	3157	11 422
Edge weight slope	2.43	2.24	2.15	2.06	2.32	1.98
L	3.96	3.15	3.02	3.23	3.56	3.16
Lr	2.46	3.28	3.07	2.80	2.63	2.72
Diameter	16	12	11	12	18	13
C%	10.82	27.71	34.81	26.69	18.42	33.86
Cr%	0.03	0.23	0.15	0.08	0.05	0.08

In addition, each network has a largest strongly connected subnetwork that has most of nodes of the network. The numbers of nodes in the largest strongly connected subnetworks are ranked as

$$11661(C) < 14599(E) < 19053(S) < 19472(F) < 30490(R) < 36412(A), \tag{2}$$

and they have the following percentage to the numbers of nodes in their corresponding whole networks:

$$82.3\%(R) < 83.7\%(C) < 85.4\%(S) < 86.4\%(A) < 87.1\%(F) < 87.6\%(E).$$

It is shown that the percentages for English is the biggest, followed by French, while those of Russian and Chinese are considerably smaller. Anyway, all the proportions of the sizes of the largest strongly connected subnetworks are very large to those of their corresponding whole networks. Therefore, the statistical parameters of each whole network are nearly the same as those of its largest strongly connected subnetwork.

3.2. Article length and number of nodes

3.2.1. Article length

Since these six networks are constructed from those articles that describe the same events, we can analyze the article lengths to find some characteristics of these languages. The length of each concatenated article means the total number of words (including repeated ones) appearing in the text. They are ordered as

$$489841(C) < 533268(A) < 577896(R) < 617948(E) < 728315(F) < 756856(S). \tag{3}$$

These data shows that in describing a same event, Chinese language uses the least number of words, followed by Arabic, while Spanish and French languages use considerably more words. Since the words are nodes of the constructed networks, the sizes of the total numbers of all the words used to describe a same event can characterize the complexity of the different languages to a certain extent. Therefore, from this point, the expression of Chinese language is the briefest and then Arabic, while expressions of Spanish and French languages are more cumbersome.

3.2.2. Number of nodes

Number of nodes in a network denotes the amount of different words in the corresponding text, and is also called the size of the network. The numbers of nodes in these six networks are ranked as

$$13927(C) < 16659(E) < 22316(S) < 22348(F) < 37044(R) < 42154(A), \tag{4}$$

which shows that the size of the Chinese network is the smallest, followed by the English network, while the Arabic and Russian networks are relatively larger.

We know that Russian language is one of languages that retain much ancient morphological changes. Grammatical relations and functions of Russian words are represented mainly by morphology and parts of speech. There are some special rules in Arabic word formations. Most of Arabic words have changes in internal roots and derivation relationships, and usually a root can derive numbers of verbs and nouns. Therefore, the main reason why the Russian and Arabic networks have more nodes is that their words have many deformations; that is, many different words are derived from a word. In addition, note that the number of nodes in the Chinese network is the smallest. It has been known that the Chinese article length is the shortest by (3). So Chinese language has the briefest expression and incline to use the least number of words to describe a same complex information as those in other five languages. This is because that Chinese words have no morphological changes.

3.3. Number of edges and edge weight with its distribution

3.3.1. Number of edges

The numbers of edges in the whole networks and their largest strongly connected subnetworks are ordered as, respectively,

$$148379(S) < 157453(C) < 159946(E) < 163010(F) < 218557(A) < 219503(R), \tag{5}$$

$$145010(S) < 155000(C) < 157597(E) < 159881(F) < 212083(A) < 212513(R). \tag{6}$$

Obviously, these two ranks for the six languages are completely same. The number of edges of the Russian network is the biggest, followed by the Arabic network, while the Spanish network is the least and the Chinese network is the second least.

3.3.2. Edge weight

In order to further study characteristics of edges, we establish the weighted networks, in which the weight of each edge is equal to the number of repetitions of the edge in the concatenated text. Moreover, for each largest strongly connected subnetwork, we compute the total number of edge weights, the average of edge weights, the biggest edge weight, and the distribution of edge weights.

The total numbers of edge weights, the averages of edge weights, the biggest edge weights of these six largest strongly connected subnetworks are ranked as, respectively,

$$427133(C) < 472417(A) < 508543(R) < 565300(E) < 669933(F) < 700906(S), \tag{7}$$

$$1.76(A) < 2.39(R) < 2.76(C) < 3.59(E) < 4.19(F) < 4.83(S), \tag{8}$$

$$664(C) < 2302(A) < 3157(R) < 9079(E) < 9487(F) < 11422(S). \tag{9}$$

As a common sense, the rank of the total numbers of edge weights is the same as that of the article lengths (see (3)) and their relative proportions are also similar.

In comparing the averages of edge weights, Arabic is the least, followed by Russian, while Spanish and French are considerably much bigger. Note that the Russian and Arabic networks have the most numbers of edges (see (5) and (6)) and the most numbers of nodes (see (4)). However, the total numbers of edge weights of their largest strongly connected subnetworks are relatively smaller. This leads to smaller averages of edge weights. This indicates that the connections among Russian and Arabic words are sparse. In addition, the Spanish network has the least number of edges (see (5)) and the biggest total number of edge weights, which results in the biggest average of edge weights. Hence, there are quite many pairs of words repeatedly many times in the Spanish text. This implies that among Spanish words are there some stringent connection ways. This further indicates to some extent that the Spanish grammar is more rigorous.

For the biggest edge weight, the Chinese network has the least one, followed by the Arabic network, while the Spanish and French networks have relatively bigger ones. It is found that the nodes with the biggest edge weight are "这", "一" in Chinese, and "de" and "la" in Spanish that are corresponding to the English words "of" and "the".

3.3.3. Distributions of edge weights

Fig. 2 plots the distributions of edge weights in the log–log scale for the six networks. Obviously, they all fit a power-law distribution, and their power-law exponents are ranked as

$$1.98(S) < 2.06(F) < 2.15(E) < 2.24(C) < 2.32(R) < 2.43(A). \tag{10}$$

In Section 3.3.2, it has been found that the biggest edge weight of the Chinese network is extraordinarily small compared with the other languages (see (9)). Note that its total number of edge weights is also the smallest (see (7)), while its average of edge weights is almost equal to the average of those of the six networks (see (8)). By analysis, we find that the main reason for the tremendous difference about the biggest edge weight is that in the other languages, a few words are needed to connect together to express a meaning of a singleton Chinese word in general.

In addition, it is found that a language owning a smaller (bigger) average of edge weights often has a bigger (smaller) distribution slope of edge weights.

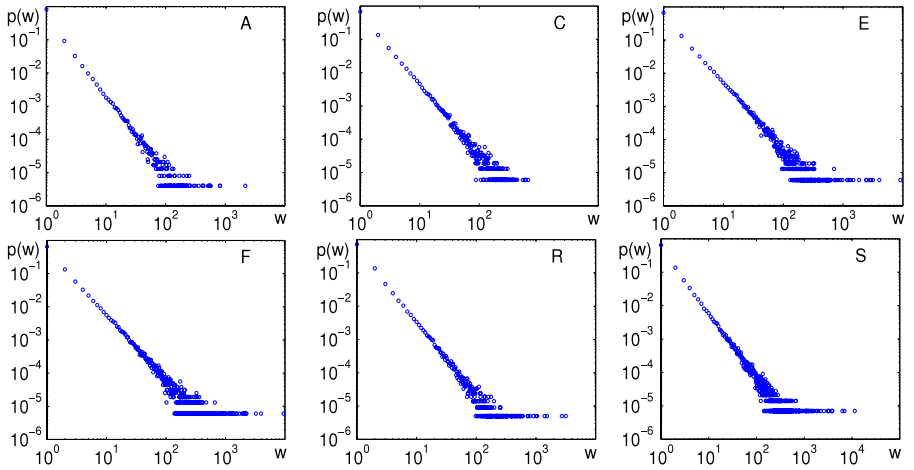

Fig. 2. Distributions of edge weights of the largest strongly connected subnetworks of the six networks in the log–log scale.

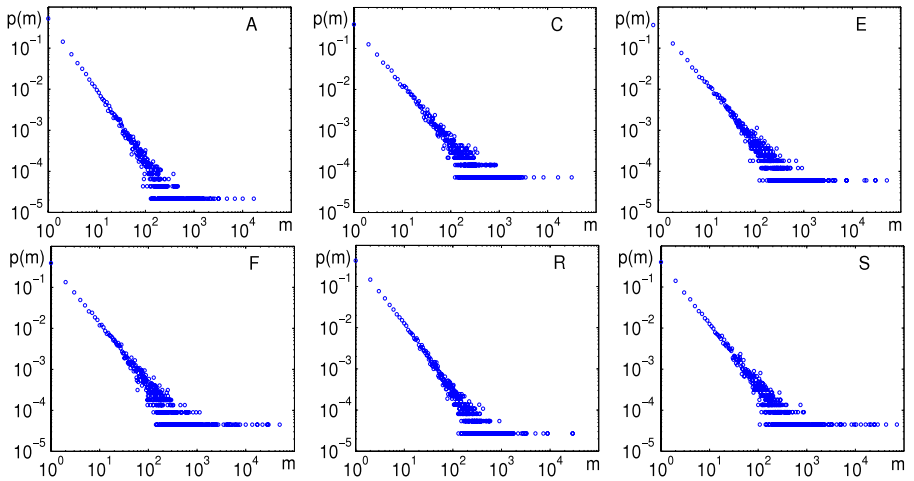

Fig. 3. Distributions of numbers of repetitions of words for the six languages in the log–log scale.

3.4. Numbers of repetitions and their distribution

The number of repetitions of a word denotes the total number of the frequency of the word appearing in the corresponding concatenated text. The averages of numbers of repetitions for these six languages are ranked as

$$10.28(A) < 15.60(R) < 32.59(F) < 33.92(S) < 35.17(C) < 37.09(E), \tag{11}$$

and the biggest number of repetitions are

$$16994(A) < 29389(R) < 31396(C) < 49848(F) < 51551(E) < 71497(S). \tag{12}$$

The words with the biggest numbers of repetitions are "في", "的", "the", "de", "B" and "de" in the Arabic, Chinese, English, French, Russian and Spanish texts, respectively.

We plot the distributions of numbers of repetitions of these six concatenated texts in the log–log scale (see Fig. 3). They present a good power-law distributions. The corresponding power-law exponent is called its repetition slope for convenience. Their repetition slopes are ordered as

$$1.47(C) < 1.50(E) < 1.56(S) < 1.61(F) < 1.73(R) < 1.77(A). \tag{13}$$

It shows that the Chinese repetition slope is the least, followed by the English repetition slope, while the Arabic and Russian repetition slopes are considerably bigger.

From the above data, one can see that both the average of numbers of repetitions and the biggest number of repetitions of Arabic and Russian are quite smaller, while their numbers of nodes are larger (see (4)). The main reason is that one inclines to use a variety of changes of words to describe an event in these two languages,. This leads to a rich usage of words, and then the number of repetitions of each word is relatively smaller in general. However, Chinese and English have the reverse

characteristic: their numbers of nodes are smaller (see (4)) and their averages of numbers of repetitions of words are much bigger.

In addition, it is again found that a network owning a smaller (bigger) average of numbers of repetitions often has a bigger (smaller) repetition slope.

3.5. In-degree, out-degree, total degree, and their distributions

For each largest strongly connected subnetwork, we compute in-degree, out-degree, total degree of each node in it, and their distributions. We now analyze them separately.

3.5.1. In-degree

The in-degree of a node means the total number of those edges whose end point is the node in a directed graph. According to Table 2, the averages of in-degrees and the biggest in-degrees for the six largest strongly connected subnetworks are ordered as, respectively,

$$5.8245(A) < 6.9699(R) < 7.6109(S) < 8.2108(F) < 10.7951(E) < 13.2922(C), \tag{14}$$

$$3776(C) < 4381(E) < 5351(F) < 5710(S) < 5732(A) < 6468(R). \tag{15}$$

The above data shows that the Chinese network has the maximal average of in-degrees and minimal biggest in-degree, followed by the English network, while the Russian and Arabic networks have relatively smaller averages of in-degree and much larger biggest in-degrees.

Remark. Surprisingly, the two ranks (14) and (15) are almost opposite. We have not yet found out what causes this phenomenon.

3.5.2. Out-degree

The out-degree of a node means the total number of those edges whose start point is the node in a directed graph. Obviously, the average of out-degree is equal to the average of in-degree for each directed network. The biggest out-degrees of the six largest strongly connected subnetworks are ordered as

$$3070(A) < 3573(F) < 3743(C) < 3913(S) < 4646(E) < 8201(R).$$

So the Arabic network has the minimal biggest out-degree, while the Russian network has the maximal biggest out-degree.

3.5.3. Total degree

The total degree of a node is equal to the sum of the in-degree and out-degree of the node. So the average of total degrees is double the average of in-degrees or out-degrees. The averages of total degrees for the six largest strongly connected subnetworks fall in the range 11.65–26.58. The rank of the biggest total degrees for the six largest strongly connected subnetworks is

$$7519(C) < 8802(A) < 8924(F) < 9027(E) < 9460(S) < 14669(R).$$

Thus, the Chinese network has the minimal biggest total degree, while the Russian network has the maximal biggest total degree. In particular, the biggest degree of the node in the Russian network is very big and about two times of that in the Chinese network.

3.5.4. Distributions of degrees

We plot the distributions of in-degrees, out-degrees, and total degrees for the six largest strongly connected subnetworks in the log–log scale (see Fig. 4). Obviously, they are power-law. The power-law exponents of the in-degree, out-degree and total degree distributions are ranked as, respectively,

$$1.58(C) < 1.81(E) < 1.89(R) < 1.92(A) < 2.00(S) < 2.09(F), \tag{16}$$

$$1.56(C) < 1.72(E) < 1.82(S) < 1.83(F) < 1.90(A) < 1.93(R), \tag{17}$$

$$1.52(C) < 1.63(E) < 1.83(A) = 1.83(F) < 1.85(S) < 1.86(R). \tag{18}$$

The power-law exponent of in-degree is bigger than that of out-degree for each language network except the Russian network, and bigger than that of total degree for all the six language networks. In addition, the power-law exponent of out-degree is bigger than or equal to that of total degree for each language network except the Spanish network. For all these three power-law exponents of degrees, the Chinese network always has the least ones, and is followed by the English network. The Russian network has the biggest power-law exponents of out-degrees and total degrees. The French network has the biggest power-law exponent of in-degrees.

It is well known that the power-law exponent is about 3 in the BA scale-free network [2], where the scale-free feature is a result of growth and preferential attachment. However, for these six language networks constructed here, all the 18 power-law exponents are less than 2.1. Therefore, there may be some other link attachment processes contributing to the scale-free

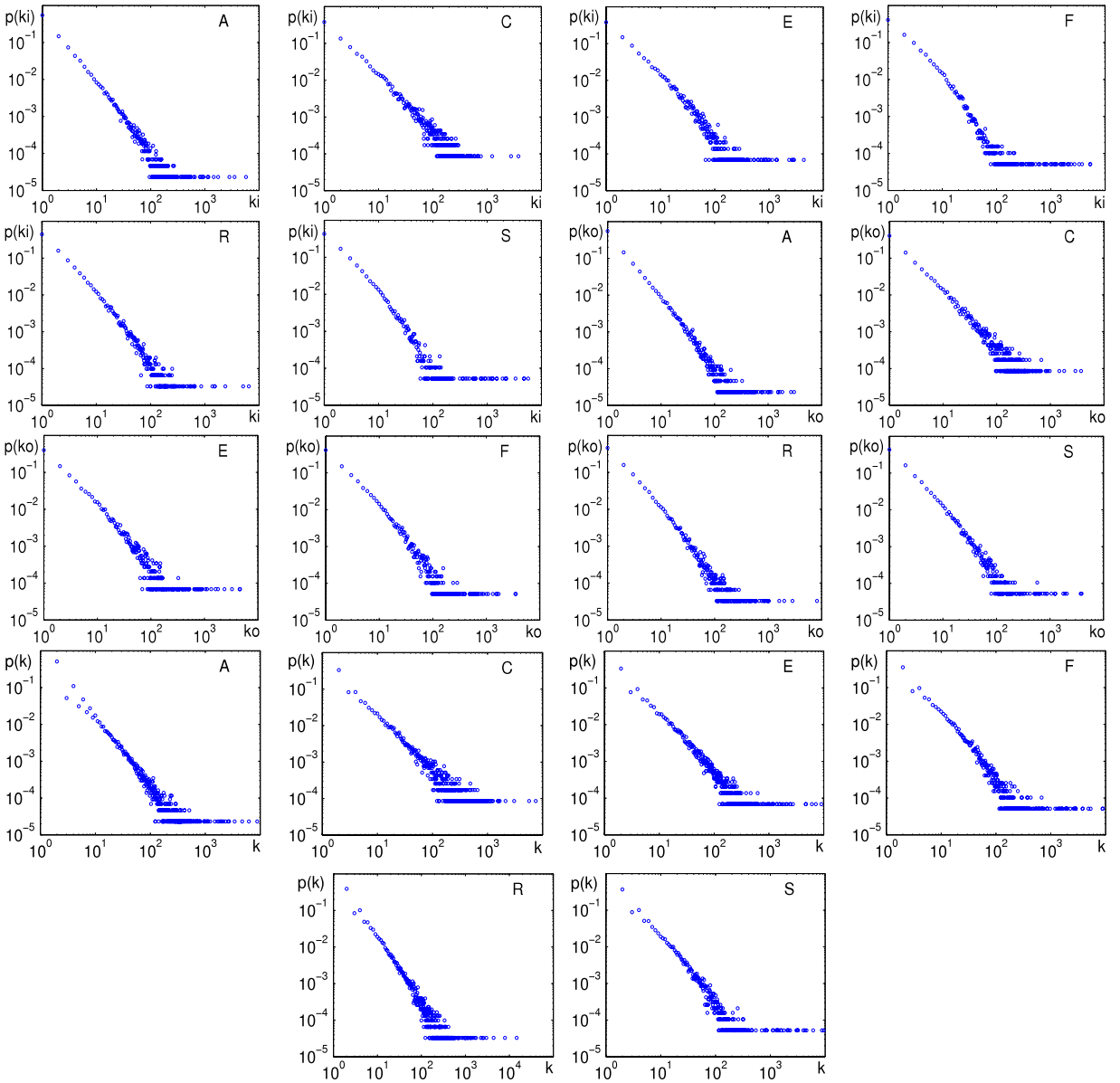

Fig. 4. Distributions of in-degrees, out-degrees, and total degrees of the largest strongly connected subnetworks of the six networks in the log–log scale.

feature, e.g., random attachment, apart from preferential attachment. A power-law degree distribution of a network implies that there are relatively a few number of nodes having very big number of connections, which are often called hubs, while most of nodes have a few connections in the network.

According to the statistics observed in the networks constructed, the nodes with the first three biggest degrees are " فسی ", " من ", " علسی " in the Arabic network; " 的 ", " 和 ", " 在 " in the Chinese network; "and", "the", "of" in the English network; "de", "des", "*à*" in the French network; "B", "Ha", "c" in the Russian network; and "y", "en", "a" in the Spanish network. Obviously, they are functional words. In addition, those words that are closely related to the topics of the articles are highly connected nodes. Most of them do not impact understanding the sentences. However, the sentences will become fragmented if these words are removed. This reflects an important characteristic of a scale-free network: the network is still connected well if a few nodes are randomly removed, but if several hubs are removed, then the entire network will break up into a number of isolated subnetworks [33].

Based on the above discussions, we have found that the Chinese network has the largest average of degrees and smallest biggest degree, and its power-law exponents of the three degree distributions are the least. This shows that the connections among Chinese words are more uniformly distributed than those of the other five language networks. This is consistent with the results "the average and distribution of edge weights of the Chinese network are at the middle of the six language

networks" obtained in Section 3.3 and the result "the Chinese text has a bigger average of numbers of repetitions" obtained in Section 3.4. Thus, it can be derived that the connection among Chinese words is relatively less dependent on those specific key words (i.e., high degree nodes). A related problem was studied in Ref. [34].

In addition, it is once more found that a network owning a smaller (bigger) average of degrees often has a bigger (smaller) power-law exponent.

3.6. Small-world effect

In this subsection, we analyze the small-world effect of our networks. Since the small-world feature is defined for connected networks. So we only consider the average shortest path lengths L, diameters D, and clustering coefficients C of the largest strongly connected subnetworks, which are ranked as, respectively,

$$3.02(E) < 3.15(C) < 3.16(S) < 3.23(F) < 3.56(R) < 3.96(A), \tag{19}$$

$$11(E) < 12(C) = 12(F) < 13(S) < 16(A) < 18(R), \tag{20}$$

$$10.82\%(A) < 18.42\%(R) < 26.69\%(F) < 27.71\%(C) < 33.86\%(S) < 34.81\%(E). \tag{21}$$

Based on the above data, all the diameters D are less than or equal to 18, and the average shortest path lengths L are all less that 4. This means that for any given two nodes in each network, it needs fewer than 4 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 words are stored in a human brain, each word can reaches another word through at most 3 intermediate words on average. This characteristic assures one of the most important functions of a human language, namely, the high speed of expressions during speech production, which makes people communicate each other by a language more convenient and fast.

By the data in Table 2, L is very close to L_r for each network, where L_r is the average shortest path length of the corresponding random network. The clustering coefficient of each network is much larger than that of the corresponding random network, i.e., $C \gg C_r$. 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 the small-world feature. Furthermore, the English network has the smallest L and the biggest C. This shows that the connections among the nodes in the English network are denser. Hence, the expressions in English are more flexible and powerful in this sense.

4. Discussions on similarities among the six languages

From the analysis of statistical parameters in Section 3, we have found that French and Spanish, Russian and Arabic, and Chinese and English are adjacent to each other in the most sequences of statistical parameters, respectively. Specifically, the values of the parameters of the French and Spanish networks are quite approximate in (1)–(4), (7)–(11) and (13)–(20). This may be related to the fact that, as we mentioned in the introduction, both Spanish and French belong to the Western Romance branch of the Romance languages family of the Indo-European language system. Although English and Russian also belong to the Indo-European language system, but they belong to different language branches. Note that most of statistical parameters of the English and Russian networks are quite different. This is not like the close relation that the statistical parameters of the Spanish and French networks have shown. This might suggest that languages belonging to a same language branch or family may share more similarities, and languages belonging to the same language system but to different branches or families may have more obvious differences. The similar results were obtained in [30–32] based on some different level language networks.

Russian belongs to the East Slavic language branch of Slavic family of the Indo-European language system, while Arabic belongs to the Semitic family of the Semitic language system. They are not in a same language system. However, our statistical results have shown that these two languages have many close values of statistical parameters (see (1)–(17) and (19)–(21)).

Chinese and English also belong to different language systems. But our statistical results have indicated that they have many approximate values of statistical parameters (see (1), (2), (4)–(6), (8), (10), (11) and (13)–(20)). Note that all the selected 100 reports were written during 2009–2012. Since the opening up of China, China's technological and cultural exchanges with the world have been more and more, and deeper and deeper. So Chinese language and culture have been certainly impacted by other languages and cultures. As the most rapid spreading international language, English language might have influenced Chinese language more deeply. This impact has attracted some Chinese linguists' attention. Perhaps the similarity of some characteristics between these two languages has a close relationship with this impact. We shall study this question in our near future research.

5. Conclusions

100 articles are selected from the UN secretary-general reports, each of which was written in Arabic, Chinese, English, French, Russian and Spanish languages, separately. By combining 100 articles in a same language into a concatenated text, 6 weighted and directed word co-occurrence networks are constructed. Their statistical characteristics are studied, including

article length; numbers of strongly connected subnetworks, nodes, edges, edge weights, repetitions and degrees (in-degree, out-degree and total degree); distributions of edge weights, repetitions and degrees; average shortest path length; diameter; and clustering coefficient.

Besides all the networks constructed exhibit scale-free and small-world features, we have obtained the following new findings: (1) The connections among English words are denser, and the expression of English language is more flexible and powerful; (2) Many parameters values of French and Spanish language networks are very approximate. This shows that these two languages share a lot of commonalities; (3) Many words in Arabic and Russian languages have much more changes. This results in both rich types of words and quite sparse connections among words; (4) Connections among Chinese words are more uniformly distributed, and the total number of different words in the text is the least. So one incline to use the least number of Chinese words to express a same complex information as those in the other five languages. This shows that its expression is concise.

It has been found that languages belonging to a same language branch or family may share many similar characteristics, while languages belonging to a same language system but to different branches or families may have some obvious differences. However, It has been found that although Chinese and English languages belong to the different language systems, but they share many similar characteristics. The same conclusion was obtained by clustering analysis in Refs. [28,32,35]. In addition, it has been observed that for all the distributions, if a parameter has a smaller (bigger) average value, then its distribution often has a bigger (smaller) slope. These findings need to be more further and deeply studied in the future research.

Acknowledgments

This research was supported by the RFDP of Higher Education of China (Grant 20100131110024), the NNSF of China (Grant 11071143), the NNSF of Shandong Province (Grant ZR2011AM002), and the RFDP of Henan Polytechnic University (Grant B2011-032).

The authors would like to thank the referees for their valuable comments and suggestions.

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] D. Fell, A. Wagner, The small world of metabolism, Nat. Biotechnol. 18 (2000) 1121–1122.
[9] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mold. Biol. Evol. 18 (2001) 1283–1292.
[10] M.E.J. Newman, The structure of scientific collaboration-networks, Proc. Natl. Acad. Sci. USA 98 (2001) 404–409.
[11] J. Cummings, R. Cross, Structural properties of work groups and their consequences for performance, Social Networks 25 (2003) 197–210.
[12] M. Kurant, P. Thiran, Layered complex networks, Phys. Rev. Lett. 96 (2006) 138701.
[13] Y. Chen, N. Li, D. He, A study on some urban bus transport networks, Physica A 376 (2007) 747–754.
[14] L. Steels, Language as a Complex Adaptive System, in: Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2000.
[15] R. Ferrer i Cancho, R.V. Solé, The small world of human language, Proc. R. Soc. Lond. Ser. B. 268 (2001) 2261–2265.
[16] M. Markosŏvá, P. Náther, Language as a small world network, in: Proceedings of the Sixth Internetional Conference on Hybrid Intelligent Systems, 2006.
[17] Y. Shi, W. Liang, J. Liu, Chi K. Tse, Structural equivalence between co-occurrences of characters and morphemes in the Chinese language, in: 2008 International Symposium on Nonlinear Theory and Its Applications, Budapest, Hungary, 7–10 Sept., 2008, pp. 94–97.
[18] W. Liang, Y. Shi, C.K. Tse, J. Liu, Y. Wang, X. Cui, Comparison of co-occurrence networks of the Chinese and English languages, Physica A 388 (23) (2009) 4901–4909.
[19] W. Liang, Y. Shi, Chi K. Tse, Y. Wang, Study on co-occurrence character networks from Chinese essays in different periods, Sci. China F, English version 55 (2012) 2417–2427; Sci. China F, Chinese version 42 (2012) 831–842.
[20] R.V. Solé, C.B. Murtra, S. Valverde, L. Steels, Language networks: their structure, function and evolution, Complexity 15 (2010) 20–26.
[21] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) ID051915.
[22] M. Sigman, G.A. Cecchi, Global organization of the Wordnet lexicon, Proc. Natl. Acad. Sci. 99 (2002) 1742–1747.
[23] L. Tang, Y. Zhang, X. Fu, Structures of semantic networks: how do we learn semantic knowledge, J. Southern Univ. 22 (2006) 413–417.
[24] H. Liu, Statistical properties of Chinese semantic networks, Chin. Sci. Bull. 54 (2009) 2781–2785.
[25] A.E. Motter, A.P.S. de Moura, Y.-C. Lai, P. Dasgupta, Topology of the conceptual network of language, Phys. Rev. E 65 (2002) ID065102(R).
[26] W.S.-Y. Wang, Language is a complex adaptive system, J. Tsinghua Univ. (Philosophy and Social Sciences) 21 (6) (2006) 5–13.
[27] B.F. Grimes, Ethnologue: Languages of the World, fourteenth ed., Summer Institute of Linguistics, Dallas, TX, 2000.
[28] H. Liu, W. Li, Language clusters based on linguistic complex networks, Chin. Sci. Bull. 55 (2010) 3458–3465.
[29] L. Sheng, C. Li, English and Chinese languages as weighted complex networks, Physica A 388 (2009) 2561–2570.
[30] O. Abramov, A. Mehler, Automatic language classification by means of syntactic dependency networks, J. Quant. Linguist. 18 (2011) 291–336.
[31] H. Liu, C. Xu, Can syntactic networks indicate morphological complexity of a language? Europhys. Lett. 93 (2011) 28005.
[32] H. Liu, J. Cong, Language clustering with word co-occurrence networks based on parallel texts, Chin. Sci. Bull. 58 (2013) 1139–1144.
[33] R. Albert, H. Jeong, A.-L. Barabsi, Error and attack tolerance of complex networks, Nature 406 (2000) 378–382.
[34] X. Chen, H. Liu, Central nodes of the Chinese syntactic networks, Chin. Sci. Bull. 56 (2011) 735–740 (in Chinese).
[35] H. Liu, Dependency direction as a means of word-order typology: a method based on dependency treebanks, Lingua 120 (2010) 1567–1578.