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.