Physica A 388 (2009) 2555–2560

Contents lists available at ScienceDirect

Physica A

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

A network of two-Chinese-character compound words in the Japanese language

Ken Yamamoto*, Yoshihiro Yamazaki

Department of Physics, Waseda University, Tokyo, 169-8555, Japan

ARTICLE INFO

Article history:
Received 14 October 2008
Received in revised form 2 December 2008
Available online 5 March 2009

Keywords:
Complex network
Japanese language
Chinese character
Small world
Scale free

ABSTRACT

Some statistical properties of a network of two-Chinese-character compound words in the Japanese language are reported. In this network, a node represents a Chinese character and an edge represents a two-Chinese-character compound word. It is found that this network has properties of being “small-world” and “scale-free”. A network formed by only Chinese characters for common use (*joyo-kanji* in Japanese), which is regarded as a subclass of the original network, also has the small-world property. However, a degree distribution of the network exhibits no clear power law. In order to reproduce the disappearance of the power-law property, a model for a selecting process of the Chinese characters for common use is proposed.

1. Introduction

It has been found that a great variety of systems, such as the internet [1,2], collaboration in science [3,4], and the food web [5,6], have network structures; the systems consist of a group of nodes which interact mutually through edges. Network science supplies some methods to understand the topological structures of such systems. Recently, it has been proved that small-world [7,8] and scale-free [9] properties are important and that many networks share these properties. For typical examples, human languages have been modeled in the framework of complex networks so as to investigate graphemic [10], phonetic [11], syntactic [12] and semantic [13] structures.

Chinese characters are main elements in the writing system of the Japanese language. One of the most remarkable features of Chinese characters is that they are ideograms; that is, a single Chinese character can convey its own meaning.

The Japanese language possesses many words constructed by combining two Chinese characters. Such words are called ‘two-Chinese-character compound words’ (*niji-jukugo* in Japanese), and we adopt the name ‘two-character compounds’ hereafter. For instance, in the Japanese-language dictionary *Kojien* [14], about 90,000 words of about 200,000 headwords are two-character compounds. So far, research on two-character compounds in the Japanese language has been concentrated mostly on morphological structures [15,16] and cognitive processes [17,18]. However, studies of the two-character compounds in the Japanese language based on network science seem to be insufficient. In the present paper, we report the analysis results of networks of two-character compounds in the Japanese language.

2. Method

First, we extracted networks of two-character compounds from the following Japanese-language dictionaries: *Kojien*, *Iwanami Kokugo Jiten*, *Sanseido Kokugo Jiten*, and *Mitsumura Kokugo Gakushu Jiten* [14]. It is noted that *Kojien*, *Iwanami*, and *Sanseido* are standard dictionaries, but *Mitsumura* is a dictionary for students of elementary and junior high school. We picked out two-character compounds from the headwords of each dictionary.

* Corresponding author.
E-mail address: yamaken@toki.waseda.jp (K. Yamamoto).

0378-4371/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2009.02.032

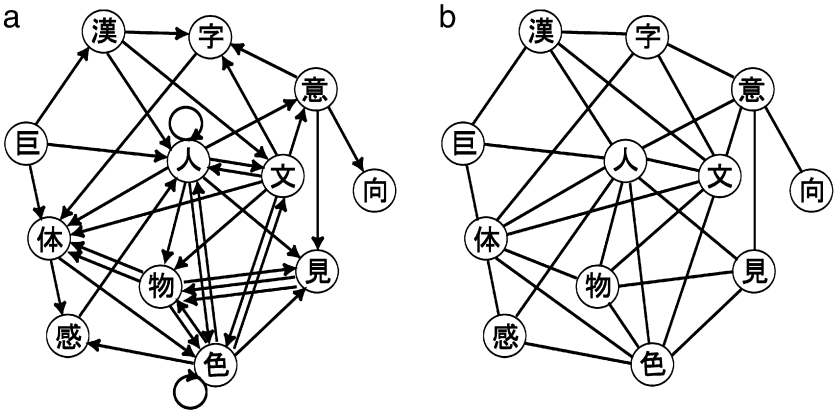

Fig. 1. A part of the network extracted from *Kojien*: (a) original network, (b) network omitting direction, multiple edges, and self loops.

Table 1
The characteristics of the maximal cluster in a network of two-character compounds. $\langle k\rangle$, ℓ, D, and C denote average degree, mean path length, diameter, and clustering coefficient, respectively. C_{rand} represents the averaged clustering coefficient of the 50 random networks of the same size in nodes and edges.

Dictionary	Nodes	Edges	$\langle k\rangle$	ℓ	D	C	C_{rand}	γ
Kojien	5458	74 617	27.3	3.14	10	0.138	0.00501	1.04
Iwanami	3904	32 150	16.5	3.31	10	0.085	0.00424	1.04
Sanseido	3444	28 358	16.5	3.32	9	0.086	0.00483	1.05
Mitsumura	1799	9 054	10.1	3.42	8	0.059	0.00255	–

In the network of two-character compounds, each Chinese character corresponds to a node, and each two-character compound formed by connecting two nodes is regarded as an edge. Each edge has a direction from an upper character to a lower character. Thus, this network is naturally viewed as a directed network with multiple edges and self loops. The direction of edges in the network deeply relates to the lexical structure and meaning of the two-character compounds. The multiplicity of edges represents the following two aspects: (i) some two-character compounds have two or more readings, and (ii) some compounds become other existing compounds when the upper and lower characters are inverted. A part of this network is depicted in Fig. 1.

In the networks we obtained, all nodes are not connective, and the whole network is made up of 169 (*Kojien*), 152 (*Iwanami*), 142 (*Sanseido*), and 8 (*Mitsumura*) clusters. In the following analysis, we consider the maximal cluster in the network of each dictionary (more than 90% of nodes belong to the maximal cluster). Since the essential features of the networks can be described even without the edge direction and multiplicity and self loops, we focus on the undirected and unweighted networks.

3. Results

The fundamental results obtained from each dictionary are summarized in Table 1. For instance, in the case of *Kojien*, a pair of two nodes is about three steps distant on average, and at most ten steps distant (see ℓ and D in this Table). The clustering coefficient C of each network is about 20 times greater than that of a random network of the same size in nodes and edges C_{rand}. Therefore, networks of two-character compounds have short path length and high clustering, as in many real networks [19]. It is found that the degree distributions of the three networks (shown in Fig. 2(a)–(c)) display the power law

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

where $p(k)$ denotes the fraction of nodes having degree k. The values of γ are nearly 1 for these three dictionaries, as shown in Table 1. However, as shown in Fig. 2(d), the degree distribution of *Mitsumura* does not exhibit a clear power-law property.

4. Restricted network formed by Chinese characters for common use

In this section, we discuss the reason why the degree distribution of *Mitsumura* does not exhibit a power law (see Fig. 2(d) for reference). There are 1945 Chinese characters designated for common use, which are called *joyo-kanji* in Japanese, selected by the Ministry of Education, Science and Culture of Japan in 1981. We call them 'common-use characters' hereafter. The common-use characters are taught during elementary and junior high school in Japan, and most Chinese characters used in Japan are these common-use characters. Moreover, Chinese characters apart from the common-use characters are