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

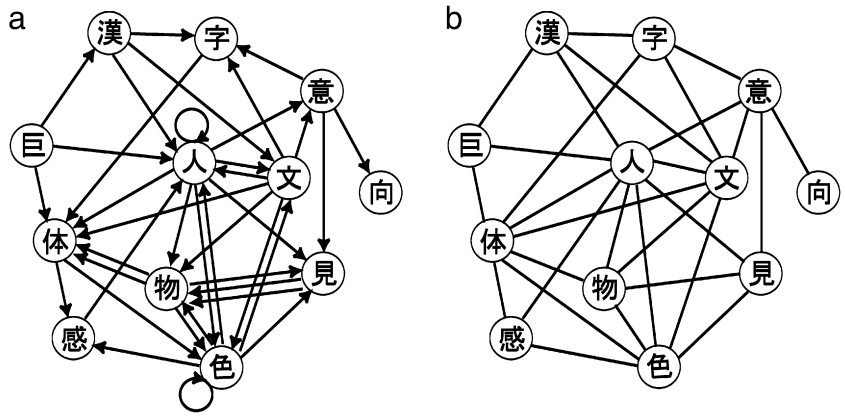

Fig. 1. A part of the network extracted from *Kojien*: (a) original network, (b) network omitting direction, multiple edges, and self loops.

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

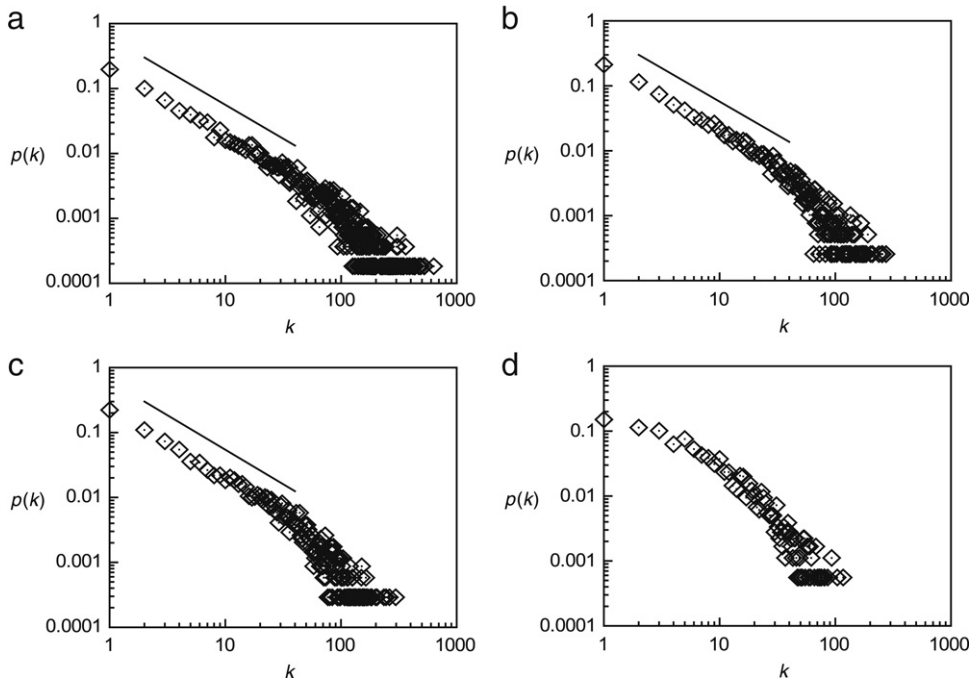

Fig. 2. Degree distribution of the network of each dictionary: (a) *Kojien*, (b) *Iwanami*, (c) *Sanseido*, and (d) *Mitsumura*. In (a)–(c), the solid lines show guidelines of power-law behaviors.

Table 2
The characteristics of the network of common-use characters.

Dictionary	Nodes	Edges	$\langle k \rangle$	ℓ	D	C
Kojien	1940	54 181	55.9	2.32	5	0.172
Iwanami	1933	26 419	27.3	2.67	6	0.111
Sanseido	1921	24 726	25.7	2.73	7	0.114
Mitsumura	1799	9 054	10.1	3.42	8	0.059

not permitted for use in legal documents. We next consider a network constructed only by the common-use characters. It is noted that this network forms a subclass of the original network.

The fundamental results of the network restricted to the common-use characters are summarized in Table 2. For the first three dictionaries in Table 2, the mean path lengths are small, and the clustering coefficients are large, compared to those presented in Table 1. On the other hand, the properties of the network of *Mitsumura* in Table 2 are the same as those in Table 1. This reflects that the two-character compounds listed in *Mitsumura* are all constructed from the common-use characters (recall that this dictionary is for students of elementary school and junior high school). As shown in Fig. 3, it is found that the degree distributions of the networks of the common-use characters do not show power-law behavior in the four dictionaries. These degree distributions share the features that there are plateaus in the range of small k ($k \lesssim 10$) and decay in large k ($k \gtrsim 10$).

5. Invasion model for selecting the common-use characters

The property of the degree distributions of the restricted networks shown above is considered to be caused by a selection process of the common-use characters. For this process, we propose a stochastic model on the 'real' maximal network of each dictionary. First, we assume that each node in the network has two states, invaded or uninvaded, and that all nodes are initially uninvaded. Then, one node is chosen randomly from the network and is turned into invaded. At each time step, one node v_i is chosen with a probability p_i from all uninvaded nodes $\{v_1, v_2, \ldots, v_n\}$ connecting to invaded nodes. The probability p_i is assumed to be given by

$$p_i = \frac{k_i^\alpha}{\sum_{j=1}^{n} k_j^\alpha} \quad (i = 1, \ldots, n), \tag{1}$$

where k_j represents a degree of a node v_j and α is a constant, which is determined below. It is noted that the case $\alpha = 0$ corresponds to random growth, that is, all v_i have equal probability of invasion, and that the case $\alpha > 0$ corresponds

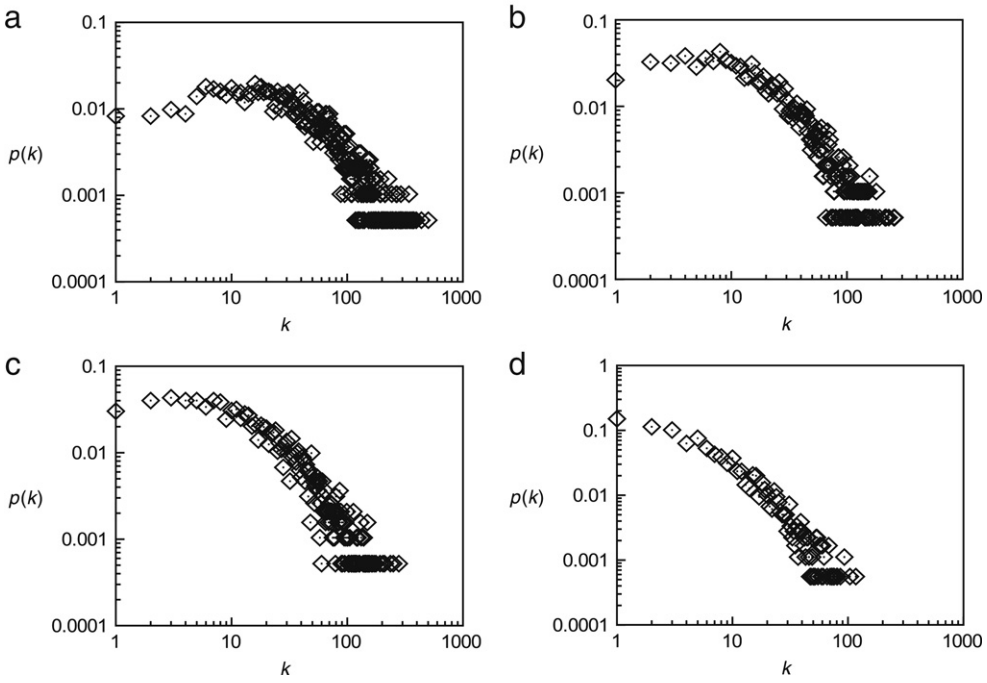

Fig. 3. Degree distributions of networks of the common-use characters: (a) *Kojien*, (b) *Iwanami*, (c) *Sanseido* and (d) *Mitsumura*.

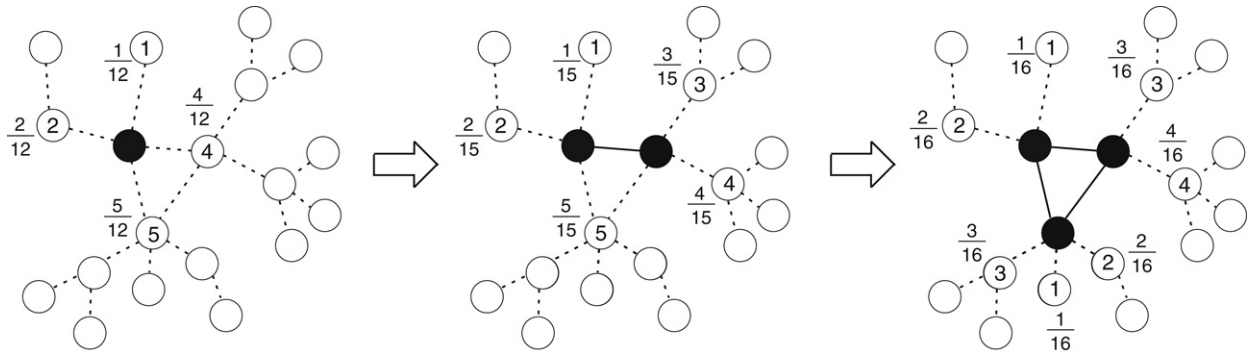

Fig. 4. An illustration of an invasion process on a network. The number on each node indicates the degree of the node, and the fractional number beside each node indicates the invasion probability ($\alpha = 1$) at each time step. The dashed and full lines indicate edges of the whole network and the generated subnetwork, respectively. Black nodes represent that they are invaded.

to 'preferential' growth, that is, a node of larger degree, invaded more easily [9,20]. The invasion process is schematically shown in Fig. 4. In this model, invaded nodes are regarded as the common-use characters. The value of α should be positive, since the common-use characters tend to have large degrees. Thus, it can be said that this model is a preferential growth process of an invaded cluster on a network, and it is similar to invasion percolation [21] or the Eden model [22].

The process of invasion was performed numerically until the number of invaded nodes amounted to the size of the network of the common-use characters in the dictionaries except *Mitsumura*. Then, we calculated $\langle k \rangle$ for the subnetwork of invaded nodes. To determine α, we require that the average degree $\langle k \rangle$ of the subnetwork of invaded nodes becomes almost the same as that of real network of common-use characters. $\langle k \rangle$ as a function of α is depicted in Fig. 5 in the range $0 < \alpha < 2$. From this figure, it is suggested that $\alpha \approx 1.3$ is appropriate for the three dictionaries.

The numerical results are in good agreement with real networks, as shown in Table 3. In addition, Fig. 6 shows that the degree distributions obtained from the numerical results are also in good agreement with those obtained from real networks.

6. Discussion

Our analysis has proved that the network of two-character compounds has both small-world and scale-free properties. The possibility of emergence of the scale-free property seems to be associated with a fitness model [23]. In the fitness model,

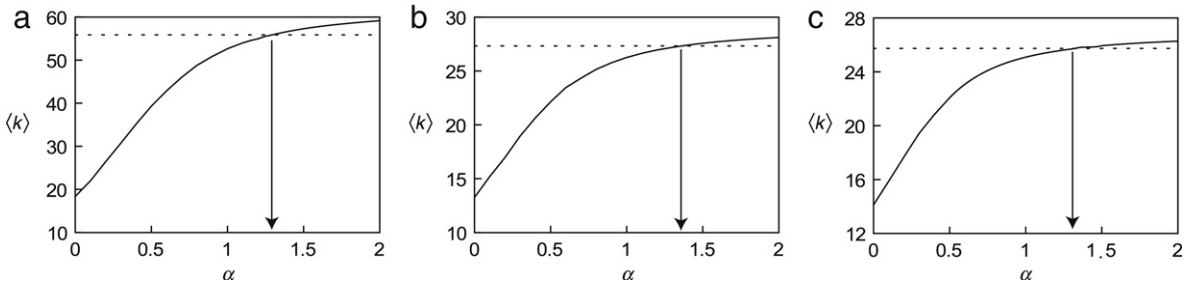

Fig. 5. Numerical results to determine the value of α. The solid lines represent the average degree $\langle k\rangle$ obtained from the model as a function of α (averaging 50 samples), and dashed lines represent $\langle k\rangle$ shown in Table 2. The intersection of solid and dashed lines indicates α: (a) $\approx$1.29 for *Kojien*, (b) $\approx$1.35 for *Iwanami*, and (c) $\approx$1.33 for *Sanseido*.

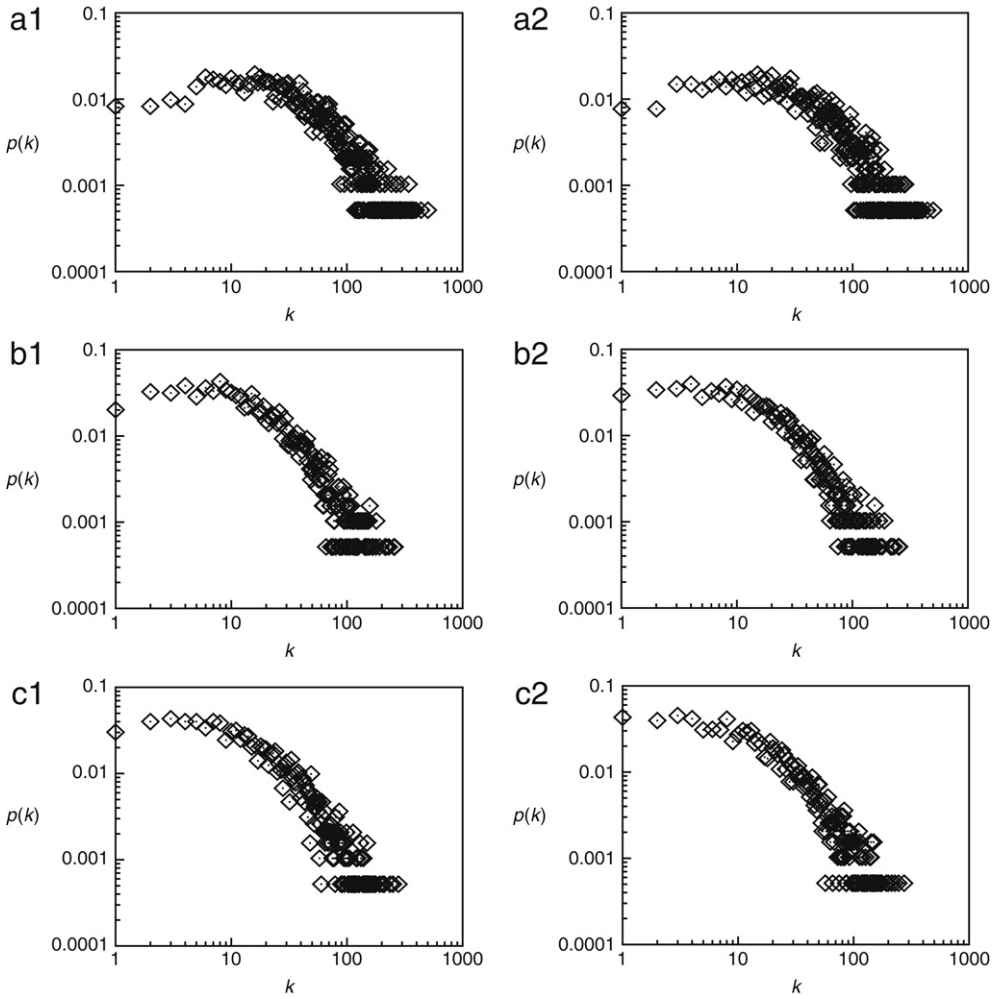

Fig. 6. Degree distributions of real common-use characters corresponding to (a1) *Kojien*, (b1) *Iwanami*, and (c1) *Sanseido*, and ones obtained from numerical results corresponding to (a2) *Kojien*, (b2) *Iwanami*, and (c2) *Sanseido*. (a1), (b1), and (c1) are identical to Fig. 3 (a)–(c).

Table 3
Comparison between real networks of common-use characters and numerical results ($\alpha = 1.3$). Numerical results are obtained by averaging 50 samples.

Dictionary	Real networks			Numerical results		
	$\langle k\rangle$	ℓ	C	$\langle k\rangle$	ℓ	C
Kojien	55.9	2.32	0.172	56.0±0.8	2.32±0.01	0.175±0.004
Iwanami	27.3	2.67	0.111	27.2±0.2	2.68±0.01	0.109±0.005
Sanseido	25.7	2.73	0.114	25.7±0.2	2.73±0.02	0.109±0.004

each node v_i in a network has a fitness x_i which is distributed independently and randomly with a given distribution function $\rho(x)$ (fitness generally represents some kind of "importance" or "sociability" of nodes). The edge between v_i and v_j is drawn with a probability given by $f(x_i, x_j)$ depending on the fitness of the nodes involved. And it is known that the fitness model can produce a power-law degree distribution.

For the network of two-character compounds, the frequency of use is uneven for each Chinese character: some characters are used quite frequently and some characters are used only in particular cases. And, it is naturally thought that the creation of two-character compounds between Chinese characters used more widely arises more frequently. Hence, there may be an effect related to the fitness model so that the network of two-character compounds has the scale-free property (a fitness in this case relates to frequency of use).

We have also found that the network of the common-use characters is connective, and that the average degree of the network is larger than that of the whole network. The invasion model proposed above is a simple method to assure connectivity and large degree of a resultant network. The model involves one parameter α, and the growth process of invaded cluster depends on the value of α. For positive α, nodes of larger degree are assigned larger invasion probability according to Eq. (1). Hence most nodes of small degree do not join a network generated from the model, and the power-law behavior in the degree distribution vanishes. Moreover, for an appropriate value of α, a plateau emerges in a range of small k in the degree distribution. It is proved that the value of α is nearly 1.3 for all three dictionaries, but we have not yet found a clear explanation for this universality.

We have confirmed that the network characteristics (Table 1) and degree distributions (Figs. 2 and 3) are essentially the same when the edge direction and multiplicity and self loops are taken into account. We think that further analysis of the direction and multiplicity will provide more precise structures of the network of two-character compounds. However, such analysis may be rather linguistic or lexical. In fact, the direction and multiplicity of edges are closely related to the individual meanings of characters and the formation principle of Japanese two-character compounds, which is classified into nine types from a grammatical point of view [24].

7. Conclusion

A network constructed by the two-character compounds in the Japanese language has short path length and high clustering (Table 1). Also, the network has a power-law degree distribution (Fig. 2), but a subnetwork restricted to the common-use characters does not show a power-law distribution (Fig. 3). The generation of the network of common-use characters can be modeled by an invasion process in which the invasion probability of nodes with degree k is proportional to k^{α} (see Eq. (1)). The exponent α is determined by consistency between the real and numerical values of $\langle k \rangle$ (Fig. 5). It confirmed that the results obtained from the model are reasonably consistent with real networks (Table 3).

References

[1] M. Faloutsos, P. Faloutsos, C. Faloutsos, Comput. Commun. Rev. 29 (1999) 251.
[2] V. Rosato, L. Issacharoff, S. Meloni, D. Caligiore, F. Tiriticco, Physica A 387 (2008) 1689.
[3] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 98 (2001) 404.
[4] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.
[5] J.A. Dunne, R.J. Williams, N.D. Martinez, Proc. Natl. Acad. Sci. USA 99 (2002) 12917.
[6] A.J. Gilbert, Ecological Indicators 9 (2009) 72.
[7] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[8] S.H. Strogatz, Nature 410 (2001) 268.
[9] A.L. Barabási, R. Albert, Science 286 (1999) 509.
[10] J. Li, J. Zhou, Physica A 380 (2007) 629.
[11] M.M. Soares, G. Corso, L.S. Lucena, Physica A 355 (2005) 678.
[12] S. Zhou, G. Hu, Z. Zhang, J. Guan, Physica A 387 (2008) 3039.
[13] M. Steyvers, J.B. Tenenbaum, Cogn. Sci. 29 (2005) 41.
[14] The sources of the Japanese-language dictionaries we used are
(a) Kojien, 4th ed., Iwanami Pub. Co., Tokyo, 1991;
(b) Iwanami Kokugo Jiten, 5th ed., Iwanami Pub. Co., Tokyo, 1992;
(c) Sanseido Kokugo Jiten, 4th ed., Sanseido Pu. Co., Tokyo, 1992;
(d) Mitsumura Kokugo Gakushu Jiten, revised ed., Mitsumura Tosho Pub. Co., Tokyo, 1991.
[15] T. Joyce, N. Ohta, Tsukuba Psychological Research 22 (1999) 45.
[16] H. Masuda, T. Joyce, Glottometrics 10 (2005) 30.
[17] A. Morita, K. Tamaoka, Brain and Language 82 (2002) 54.
[18] K. Tamaoka, Reading and Writing 18 (2005) 281.
[19] R. Albert, A.L. Barabási, Rev. Modern Phys. 74 (2002) 47.
[20] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 85 (2000) 4629.
[21] D. Wilkinson, J.F. Willemsen, J. Phys. A 16 (1983) 3365.
[22] M. Eden, in: F. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. IV, University of California, Berkeley, 1961.
[23] G. Caldarelli, A. Capocci, P.De. Los Rios, M.A. Muñoz, Phys. Rev. Lett. 89 (2002) 258702.
[24] M. Nomura, Nihongogaku 7 (1988) 44 (in Japanese).