2016 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2016, Yugawara, Japan, November 27th-30th, 2016

Analysis on Differences of Japanese and English Languages by the Complex Network Theory

Mayumi Tatara†, Yutaka Shimada†,‡, Kantaro Fujiwara†,‡ and Tohru Ikeguchi†,‡

†Department of Management Science, Graduate School of Engineering, Tokyo University of Science
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika-ku, Tokyo, Japan

Abstract—The complex network based approaches considerably enhance our understanding of many real systems, for example, the Internet, human relations and neural networks. Languages can also be analyzed by the complex network based approach, because languages are described as a network consisting of words and their adjacency relations. Even though there are several researches on the language networks, they mainly focus on a specific language, and there are few researches comparing different languages from the viewpoint of complex networks.

In this paper, we generate the language networks from literature written in Japanese and English, and investigate differences of their network structures between Japanese and English. As a result, the structural properties of Japanese language networks are clearly different from those of English ones.

1. Introduction

Many natural, social, and artificial systems are described as networks which consist of a set of links and a set of nodes. The complex network theory has revealed common structural properties underlying the networks obtained from various types of real systems [1, 2]. Languages have also been analyzed from the viewpoint of complex networks. For example, Ref. [3] shows that the language networks describing co-occurrence of words have small-world and scale-free properties. The language networks have also been used as one of benchmarks for evaluating community detection methods [4]. In these previous studies, the language networks are generated from one specific language. In this paper, we raise a question whether we can quantify differences between one language and other languages from the viewpoint of network structures. To accomplish this issue, we generate the language networks from Japanese and English literature, and investigate differences between Japanese and English languages by analyzing their network structures.

2. Data

In this paper, we used Japanese literature provided from the web site "Aozora-bunko" [5] and English literature provided from the web site "Project Gutenberg" [6]. We choose 36 literature (18 each) which have higher access rankings in these websites [5, 6]. Tables 1 and 2 show authors and titles of the Japanese literature and the English literature that we used in this paper. We generated 36 language networks from these literature.

Table 1: Authors and titles of Japanese literature

Author	Title
Kenji Miyazawa	Ginga tetsudo no yoru
Ryunosuke Akutagawa	Imogayu
Soseki Natsume	Kokoro
Osamu Dazai	Hashire Merosu
Motojiro Kajii	Lemon
Nankichi Niimi	Gongitsune
Franz Kafka (Translated by Yoshito Harada)	Henshin
Katai Tayama	Futon
Mimei Ogawa	Akai rosoku to ningyo
Torahiko Terada	Kagakusha to atama
Ohgai Mori	Takasebune
Kyoka Izumi	Koyahijiri
Kotaro Takamura	Chieko no hansei
Juza Unno	Daiuchu enseitai
Ango Sakaguchi	Mo gunbi ha iranai
Kunihiko Sugawa	Mujinto ni ikiru jurokunin
Sakutaro Hagiwara	Nekomachi
Kunio Yanagida	Yama no jinsei

3. Methods

3.1. How to generate language networks

In our study, we generated language networks by the following two methods.

Method 1 We defined nodes as words and links as the adjacency relation between the words, where each word connects with its nearest neighbors in the same sentence by the links. Figure 1(a) shows how to generate the language network by the method 1.

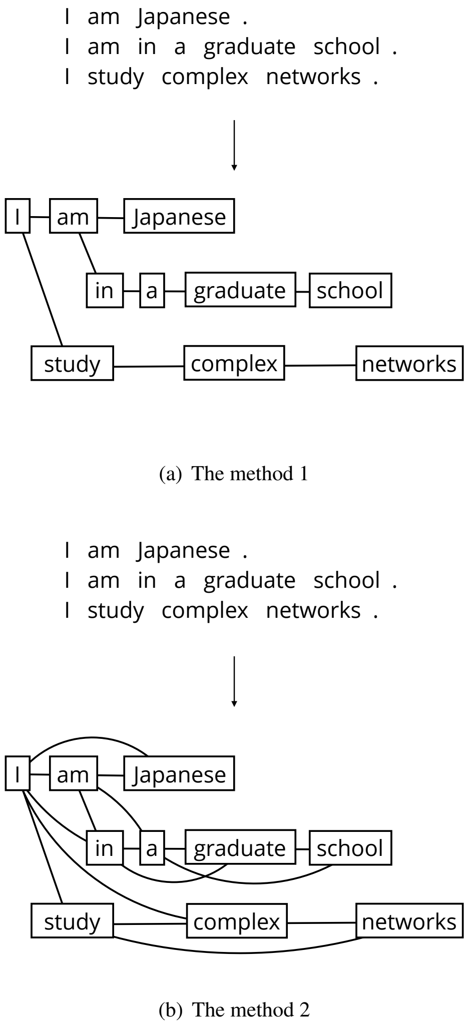
Table 2: Authors and titles of English literature

Author	Title
Lewis Carroll	Alice's Adventures in Wonderland
Mark Twain	The Adventures of Tom Sawyer
The Brothers Grimm	Grimm's Fairy Tales
J. M. Barrie	Peter Pan
Charles Dickens	A Christmas Carol in Prose; Being a Ghost Story of Christmas
	A Tale of Two Cities
Jane Austen	Pride and Prejudice
	Emma
Arthur Conan Doyle	The Adventures of Sherlock Holmes
Henrik Ibsen	A Doll's House
Jonathan Swift	A Modest Proposal
Daniel Defoe	The Life and Adventures of Robinson Crusoe
Mary Wollstonecraft Shelley	Frankenstein or The Modern Prometheus
Oscar Wilde	The Picture of Dorian Gray
Bram Stoker	Dracula
Lee Sutton	Venus Boy
Jack Sharkey	The Secret Martians
Robert Louis Stevenson	Treasure Island

Method 2 Each word connects with its next nearest neighbors in the same sentences. Figure 1(b) shows how to generate the language network by the method 2.

No links have weights and directions in this paper. Even if the same pairs of adjacent words occur more than once in the same literature, the number of links between these nodes is only one. In addition, symbols including punctuations and brackets are not contained in the language networks, and the words which are adjacent to these symbols are not connected with links. Self-loops such as "very very" are omitted from the networks.

To construct language networks for Japanese texts, we have to use a morphological analysis tool, because in Japanese texts, each word is not separated by a space. The morphological analysis enables us to automatically identify words from Japanese texts. In this paper, we used MeCab which is one of the morphological analysis tools for Japanese language [7].



(a) The method 1

(b) The method 2

Figure 1: How to generate language networks.

3.2. Measures of complex networks

We calculated the characteristic path length and the clustering coefficient of the language networks generated by the methods 1 and 2. The characteristic path length is the average of the shortest path lengths of all pairs of two nodes in the network. Let l_{ij} be the shortest path length from the node v_i to the node v_j, and N be the number of nodes in the network. The characteristic path length L is then given by

$$L = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} l_{ij}. \tag{1}$$

The clustering coefficient C is defined as follows. Let k_i be the degree of the node v_i, and then at most $k_i(k_i - 1)/2$ links can exist between the adjacent nodes of v_i. The ratio of the actual number of links between the adjacent nodes of v_i to the maximum number of such links is

$$C_i = \frac{\text{the number of links between the adjacent nodes of } v_i}{{}_{k_i}\mathrm{C}_2}. \tag{2}$$

The clustering coefficient C is then defined as

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{3}$$

By calculating these values, we compared the network structures generated from the Japanese and English literature. To compare the language networks with different sizes, L and C are normalized by L and C of randomized networks which are generated by randomizing links in the original language network so that the degree of each node are preserved [8]. In this randomization method, we first randomly selected two links which do not share nodes. Next, we selected one node from each of these links at random, and exchanged them. Repeating this procedure, we generated randomized networks.

4. Result

Table 3(a) shows the number of nodes and the number of links in the Japanese language networks, and Table 3(b) shows those in the English ones. In both Tables 3(a) and 3(b), M_1 indicates the number of links in the language network generated by the method 1, and M_2 indicates that by the method 2. From Tables 3(a) and 3(b), the number of links is about twice as large in the language networks generated from the method 1 as in those from the method 2 in both Japanese and English literature.

Figure 2 shows structural comparison between the Japanese and the English language networks by the normalized characteristic path length L_O/L_R and the normalized clustering coefficient C_O/C_R, where L_O is the characteristic path length of the original network, L_R is that of the randomized network, C_O is the clustering coefficient of the original network, and C_R is that of the randomized network. From Fig. 2, the distribution of $(C_O/C_R, L_O/L_R)$ is

Table 3: The numbers of nodes and edges in (a) Japanese language and (b) English language networks

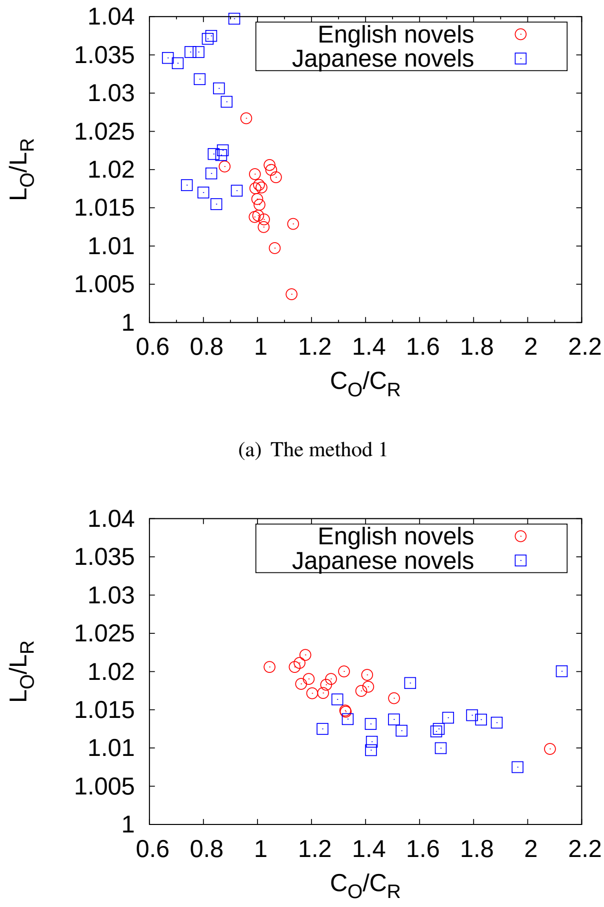
(a)

Title	N	M_1	M_2
Ginga tetsudo no yoru	2,586	9,386	18,457
Imogayu	1,854	5,135	9,781
Kokoro	6,617	30,196	61,588
Hashire Merosu	1,373	3,222	5,982
Lemon	866	1,954	3,673
Gongitsune	669	1,616	3,098
Henshin	3,220	12,531	25,184
Futon	3,844	13,281	25,841
Akai rosoku to ningyo	737	2,147	4,066
Kagakusha to atama	576	1,371	2,556
Takasebune	1,076	2,915	5,561
Koyahijiri	3,905	12,798	24,358
Chieko no hansei	1,816	4,761	9,103
Daiuchu enseitai	3,392	12,657	24,964
Mo gunbi ha iranai	1,220	3,060	5,848
Mujinto ni ikiru jurokunin	5,018	21,337	42,663
Nekomachi	1,414	3,769	7,079
Yama no jinsei	8,308	31,767	63,166

(b)

Title	N	M_1	M_2
Alice's Adventures in Wonderland	2,647	12,864	24,407
The Adventures of Tom Sawyer	7,499	35,086	65,533
Grimm's Fairy Tales	4,939	34,920	66,559
Peter Pan	4,964	23,424	43,614
A Christmas Carol in Prose; Being a Ghost Story of Christmas	4,365	16,405	30,600
A Tale of Two Cities	10,150	58,772	108,891
Pride and Prejudice	6,489	47,762	90,223
Emma	7,351	57,393	107,570
The Adventures of Sherlock Holmes	8,284	44,198	82,617
A Doll's House	2,451	11,161	20,451
A Modest Proposal	1,075	2,716	5,220
The Life and Adventures of Robinson Crusoe	6,704	44,912	84,972
Frankenstein or The Modern Prometheus	7,092	37,587	70,462
The Picture of Dorian Gray	7,075	34,451	64,304
Dracula	9,701	60,779	113,058
Venus Boy	3,475	18,424	34,198
The Secret Martians	5,775	24,083	45,063
Treasure Island	6,166	32,359	60,579

classified into two classes corresponding to the Japanese language networks and English ones. In case of the language networks generated by the method 1 (Fig. 2(a)), the distribution of the Japanese language networks is located in the upper-left part, and that of the English language networks is located in the bottom-right part. However, in the case of the method 2 (Fig. 2(b)), the distribution of the Japanese language networks is located in the bottom-right part, and that of the English language networks is located in the upper-left part. According to Table 3, when we change the method for generating networks from the method 1 to 2, the number of links is equally doubled in almost all Japanese and English language networks. In spite of this fact, the changes of L_O/L_R and C_O/C_R in the Japanese language networks are larger than those in the English ones. These differences between Japanese and English language networks might be due to the difference of grammatical features between Japanese and English languages.



(a) The method 1

(b) The method 2

Figure 2: The results of C_O/C_R and L_O/L_R for the language networks generated by (a) the method 1 and those by (b) the method 2.

5. Conclusion

In this paper, we generated the networks from 36 literature written in Japanese and English, and investigated their network structures. As a result, the characteristic path lengths and the clustering coefficients of the Japanese language networks and the English language networks are classified into different classes. In addition, distribution tendency depends on how to generate networks. These differences might come from the grammatical features of Japanese and English languages.

Acknowledgments

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant-in-Aid for Research Activity Start-up (No. 26880020) and by Grant-in-Aid for Young Scientists B (No. 16K16126) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No. 15K12137) and by Grant-in-Aid for Young Scientists B(No. 16K16138) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No. 15KT0112) from JSPS.

References

[1] Duncan J. Watts and Steven H. Strogatz, “Collective dynamics of ‘small-world’ networks,” Nature, **393**, 6684, 440–442, 1998.

[2] Albert-László Barabási and Réka Albert, “Emergence of scaling in random networks,” Science, **286**, 5439, 509–512, 1999.

[3] Ramon Ferrer i Cancho and Richard V. Solé, “The small world of human language,” Proceedings of the Royal Society of London B: Biological Sciences, **268**, 1482, 2261–2265, 2001.

[4] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” Physical Review E, **74**, 3, 036104, 2006.

[5] Aozora-bunko, http://www.aozora.gr.jp/, (accessed 2016/01/21).

[6] Project Gutenberg, http://www.gutenberg.org/, (accessed 2016/01/21).

[7] MeCab:Yet Another Part–of–Speech and Morphological Analyzer, http://taku910.github.io/mecab/, (accessed 2016/01/21).

[8] Sergei Maslov and Kim Sneppen, “Specificity and stability in topology of protein networks,” Science, **296**, 5569, 910–913, 2002.