Syntactic Dependency Structures of English and Japanese

大矢　政徳
Masanori OYA

Abstract

This study introduces the syntactic dependency structure of English and Japanese, and explores the possibility of graph-centrality measures to calculate the similarity between the syntactic dependency structure of an English sentence and that of Japanese sentence which semantically corresponds to the English sentence. The sentences in a small-size English-Japanese parallel corpus are parsed, and the parse output is used to calculate the graph-centrality measures of these sentences' syntactic dependency structures. It is expected that the measures are similar in proportion to the similarity of the syntactic dependency structures of an English sentence and a Japanese one.

Keywords : syntactic dependency structure, graph centrality, dependency parsing

キーワード： 統語依存構造、グラフ中心性、依存関係構文解析

1. Introduction

This study introduces the syntactic dependency structure of English and Japanese, and explores the possibility of graph-centrality measures to calculate the similarity between the syntactic dependency structure of an English sentence and that of Japanese sentence which semantically corresponds to the English sentence. It has been accepted as a matter of fact that English sentences have syntactic structures different from Japanese ones. However, this judgment is a subjective one without any theoretical basis. An objective measure for the similarity of syntactic structures across different languages will have both pedagogical and theoretical significance. Such measures must be both unique（one numerical value can be calculated from a given sentence）and universal（one numerical value can be calculated from a sentence across languages）. This study is an attempt to use the typed-dependency trees for English and Japanese sentences as the input to calculate their centrality measures, and to explore how these centrality measures reflect the similarity of their sentence structures. The structure of this study is as follows. Section 2 introduces the theoretical background:

おおやまさのり：外国語学部英米語学科専任講師

dependency grammar and graph-centrality measures. Section 3 reports the corpus-based experiment, and Section 4 concludes this study along with suggestions for further study.

2. Theoretical Background

2.1. Dependency Grammar

Dependency Grammar is a set of syntactic theories which focus on the dependency relationship among words in a sentence. Since it was first proposed in Tesnière (1959), Dependency Grammar has been developed by a number of researchers, e.g., Extensible Dependency Grammar by Debusmann (2003), Word Grammar by Hudson (2010), and Stanford Dependency by de Marneffe et al. (2006).

The dependency relationship among words in a sentence can be represented in a typed-dependency directed acyclic graph (henceforth DAG). For example, the sentence "I am studying graph theory" can be represented as follows:

(1) I am studying graph theory.

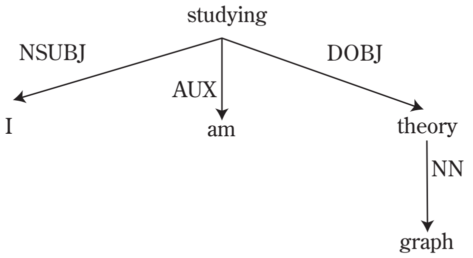

Figure 2.1 the typed-dependency tree for "I am studying graph theory."

In Figure 2.1, each of the words is represented as one node, and these nodes are connected by an arc with a label. The direction of the arc represents the direction of the dependency, and the node from which an arc starts is the head, and the node to which the arc ends is the tail. For example, the word "studying" is the head of the words "I", "am" and "theory", and the labels of these dependencies are "nsubj", "aux" and "dobj", respectively.

Consider the following sentences (2) and (3); both of them have three words, yet the dependency relationship among them, and the structural characteristics are different:

(2) Read this book.

(3) Sarah read it.

The typed-dependency DAG representations for (2) and (3) are Figure 2 and Figure 3, respectively:

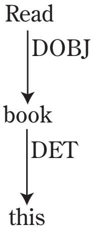

Figure 2.2 the typed-dependency DAG representation for “Read this book”.

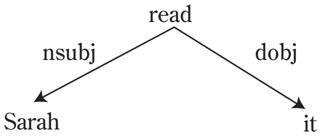

Figure 2.3 the typed-dependency DAG representation for “Sarah read it”.

The dependency relationship in the former DAG is deeper than that in the latter. There are two arcs from the root node to the terminal node in the DAG in Figure 2.2, while there is one arc from the root to the terminal node in the DAG in Figure 2.3. On the other hand, the dependency relationship in the latter DAG is wider than the former. The root is connected to two other nodes in the DAG in Figure 2.3, while the root is connected to one node in Figure 2.2. For sentences with the same word count, we can have typed-dependency DAGs with different widths and depths. The width and the depth of a given DAG can be calculated as *degree centrality* and *closeness centrality* in a well-defined manner.

2.2. Degree centrality and closeness centrality

Degree centrality is defined as the number of nodes connected to one node (Freeman, 1979). A star graph, or a graph with one node being connected to all the other nodes, has the highest degree centrality. Degree centralization is defined as the variation in the degrees of nodes divided by the maximum degree which is possible in a network of the same node count (de Nooy et al., 2005). In this study, the term degree centrality is used to denote degree centralization.

The distance from one node from another is the number of arcs between them. Closeness centrality is defined by the average distance from a given node to another in a graph (Freeman, 1979; Wasserman & Faust, 1994). In typed-dependency DAG representation, the most relevant distance is that from the root node to all the other nodes, since it represents the depth of dependency. Closeness centrality decreases in proportion to the embeddedness, or the distance from the root to the other nodes.

2.3. Typed-dependency tree centralities as similarity measures for their typed-dependency tree representations

This section deals with typed-dependency tree centralities as similarity measures for their syntactic structure. Suppose we have a string "W1 W2 W3 W4 W5". This string can be parsed to have a number of different typed-dependency trees. We can say that one extreme instance of these typed-dependency trees is the flattest typed-dependency tree (Figure 2.4), and the other extreme instance is the most embedded typed-dependency tree (Figure 2.5).

The typed-dependency tree in Figure 2.4 has a flat setting.

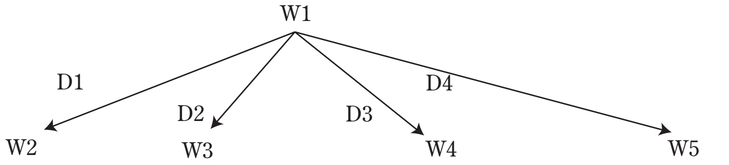

Figure 2.4 the flattest possible typed-dependency tree for a string "W1 W2 W3 W4 W5"

The typed-dependency tree in Figure 2.5 has a linear setting.

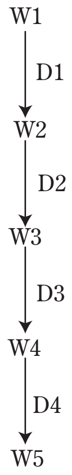

Figure 2.5 the most embedded possible typed-dependency tree for a string "W1 W2 W3 W4 W5"

These two dependency graphs (the star-graph and the line-graph) are two extreme cases of the same number of nodes (or words) in terms of centrality; the star-graph has the highest degree centrality and the highest closeness centrality (shortest *path length* (Oya 2010b) which is the inverse of closeness centrality), and the line-graph has the lowest degree centrality and the lowest closeness centrality. All the other dependency graphs of the same node number fall between them.

Oya (2010b) argued that it is possible to compare the centrality measures of graphs with different node numbers, because these centrality measures are calculated with the number of nodes in a given graph as the denominator and hence normalized; however, Oya (2012) showed that longer sentences have the tendency to have smaller degree centrality. Therefore, the argument by Oya (2010b) should be discarded and, when using degree centrality, the difference in the word count must be taken into consideration.

Centrality measures of typed-dependency trees can be employed to indicate the structural similarity among more than one typed-dependency tree of two languages. If we choose one sentence from English and another from Japanese which is a translation equivalent of the English sentence, and calculate their similarity indices, we can quantify the structural similarity between them. For example, consider the following English-Japanese sentence pairs [1, 2]:

(4)
a. Sarah has already read this book.
b. Sarahwa mou kono honwo yonda.

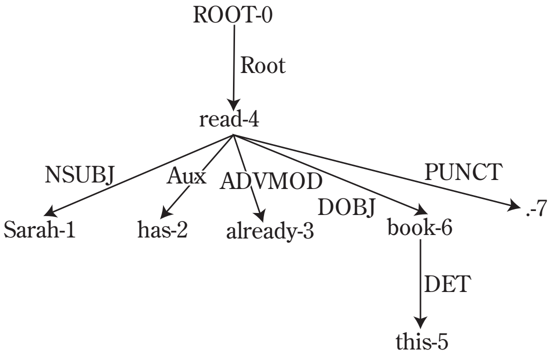

Figure 2.6 the typed-dependency tree for (4a) "Sarah has already read this book."

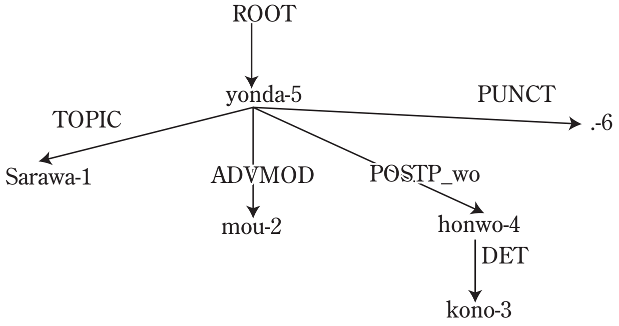

Figure 2.7 the typed-dependency tree for (4b) “Sarawa mou kono honwo yonda”.

(5)

a. The convenience store is on the other side of the street.

b. Konbiniwa toorino mukougawani arimasu.

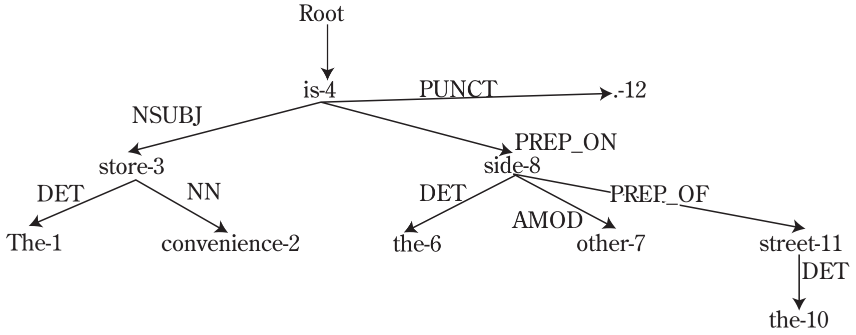

Figure 2.8 the typed-dependency tree for (5a) “The convenience store is on the other side of the street.”

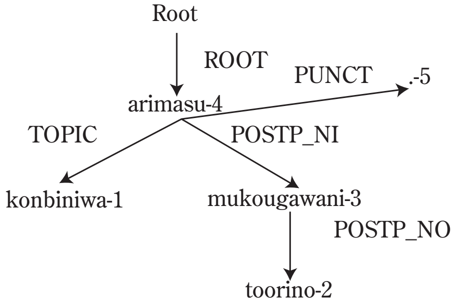

Figure 2.9 the typed-dependency tree for (5b) “konbiniwa toorino mukougawani arimasu.”

(6)

a. There seems to be something wrong with this computer.

b. Kono pasokonwa dokoka koshoushiteirumitaida.

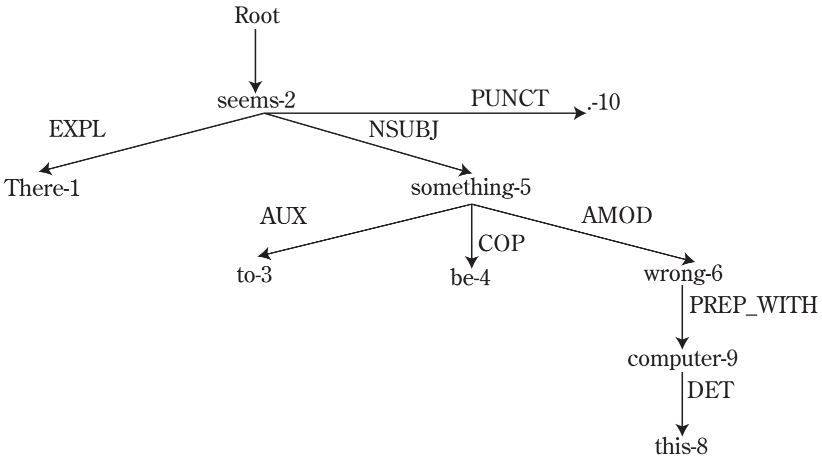

Figure 2.10 the typed-dependency tree for (6a) "There seems to be something wrong with this computer."

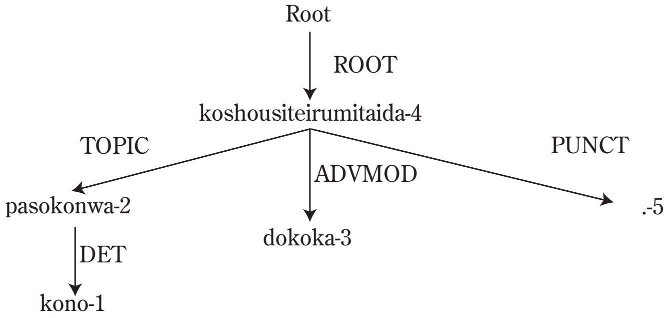

Figure 2.11 the typed-dependency tree for (6b) "kono pasokonwa dokoka koshousiteirumitaida."

The degree centralities and the closeness centralities of these English-Japanese sentence pairs are as follows:

Table 2.1 the degree centralities and closeness centralities of the typed-dependency trees

	(4a)	(4b)	(5a)	(5b)	(6a)	(6b)
D	0.81	0.77	0.24	0.7	0.29	0.7
C	0.53	0.53	0.41	0.55	0.38	0.55

D: degree centrality; C: closeness centrality

The degree centralities of English sentences (4a), (5a), and (6a) show that (4a) is flatter than other two (larger degree centralities indicate flatter typed-dependency trees). The degree centralities of Japanese sentences (4b), (5b), and (6b) show that they are similar in terms of the flatness of their typed-dependency trees. Comparing the English-Japanese pair of (4a) and (4b) to that of (5a) and (5b), (5b) is much flatter than (5a).

The reason of this difference can be subsumed to the fact that the sentences in a given English-Japanese pair have different word counts[(3)]. The word count of (4a) is 8, and that of (4b) is 7 (the root included). The word count of (5a) is 11, and that of (5b) is 6. Oya (2012) showed that sentences with smaller number of words tend to have larger degree centralities. However, only three English-Japanese sentence pairs are too small in size to draw any conclusion here. In order to understand the degree centrality and structural similarity of sentences, it is necessary to calculate the degree centralities of more English-Japanese pairs, with their word counts taken into consideration.

3. Data Analysis

3.1. Description of the data

The data chosen for the purpose introduced above is “Eisakubun Kihon 300 Sen” (Basic 300 Sentences for English Composition; henceforth Basic 300) (Iida 2010). Basic 300 contains 300 English sentences along with their Japanese translations (In some instances in Basic 300, more than one English sentence is counted as one single sentence; hence the actual number of the English sentences in Basic 300 is 339). Basic 300 is compiled for Japanese high school students to memorize basic syntactic structures of English; therefore, it contains important syntactic constructions of English.

3.2. Procedure

The English sentences in Basic 300 are parsed by Stanford Parser (de Marneffe and Manning 2008, 2011). The output option is set to Collapsed Tree, in which prepositions are collapsed to the dependency type. The parse output for each sentence is checked in terms of the dependency relation and dependency type. If any of the dependency relations and types in the output file is found incorrect, it is manually corrected. The centrality measures of the manually-corrected parse output are calculated by a Ruby script originally written for this study.

The Japanese sentences in Basic 300 are parsed by KNP ver. 4 (Kurohashi and Nagao 1992, 1994, 1998; Kawahara and Kurohashi 2007), and the parse output is converted automatically by an original Ruby script (the conversion policy is based on Oya 2010a) into the same format as that of Stanford Parser. The result is manually corrected, and the centrality measures are calculated by the same Ruby script used for English parse output.

3.3. Results

The descriptive statistics in Table 3.1 show that English sentences have smaller degree centralities than Japanese on average, which means that English sentences are less flat than Japanese ones. They have smaller closeness centralities than Japanese on average, which means that English sentences are more embedded than Japanese ones.

Table 3.1 the descriptive statistics of the degree centralities, closeness centralities, and word counts of the sentences in Basic 300 (n = 339)

	Degree		Closeness		WordCount	
	English	Japanese	English	Japanese	English	Japanese
Mean	0.39	0.53	0.43	0.50	11.04	6.61
SD	0.18	0.26	0.08	0.13	3.03	2.04

The distribution of sentences with the horizontal axis the degree centralities and with the vertical axis the word counts reveals that the variation of degree centralities which English sentences can take is wider than that of those which Japanese sentences can take, as is shown in Figure 3.1:

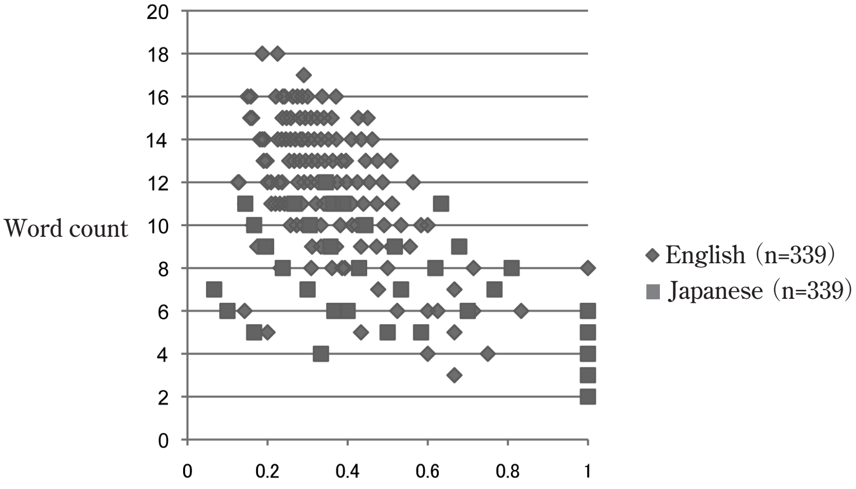

Figure 3.1 the distribution of degree centralities and word counts (n = 339)

The different distributions can be explicated if we focus on the sentences with the same word count. For example, in the small-scale corpus used here, eight-word English sentences have one of the 10 different values of degree centrality, while eight-word Japanese sentences have one of the 4 different values of degree centrality. In addition to this, the distribution of degree centralities of these Japanese eight-word sentences is concentrated on one particular value, while that of these English eight-word sentences is not:

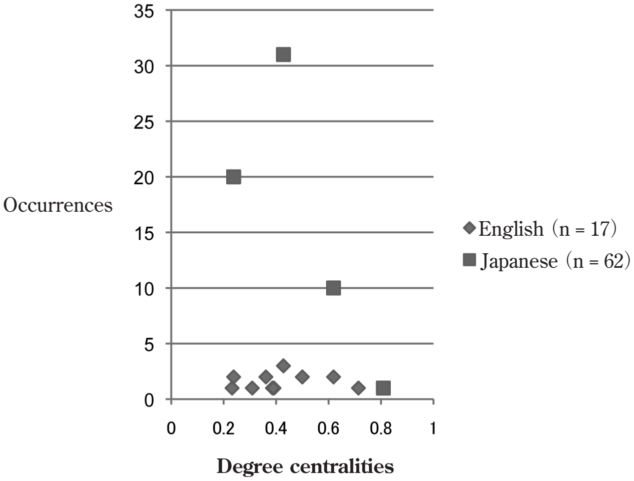

Figure 3.2 the degree centralities and the occurrence of eight-word sentences in Basic300

For nine-word sentences, English sentences have one of the 10 different values of degree centrality, while nine-word Japanese sentences have one of the 5 different values of degree centrality. Again, the distribution of degree centralities of these Japanese nine-word sentences is concentrated on one particular value. The distribution of degree centralities of these English nine-word sentences is also concentrated on one particular value, but not as strongly as in Japanese:

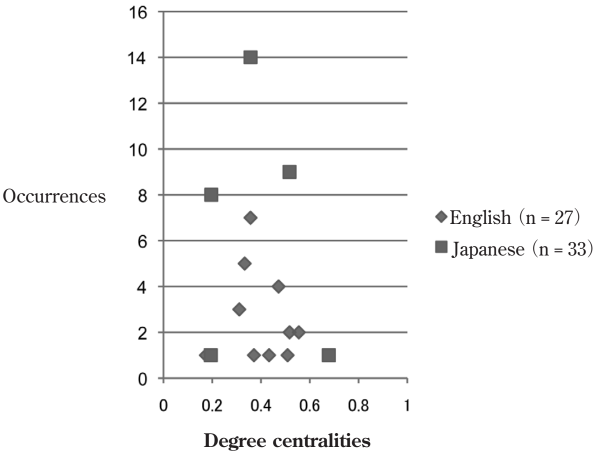

Figure 3.3 the degree centralities and the occurrence of nine-word sentences in Basic300

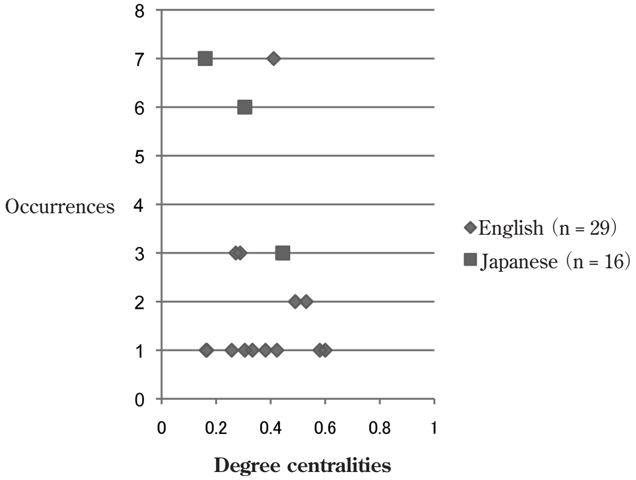

Figure 3.4 the degree centralities and the occurrence of ten-word sentences in Basic300

The distribution of closeness centralities makes a good contrast with that of degree centralities. As figure 3.5 below shows, closeness centralities decrease in proportion to the word counts in the sentences:

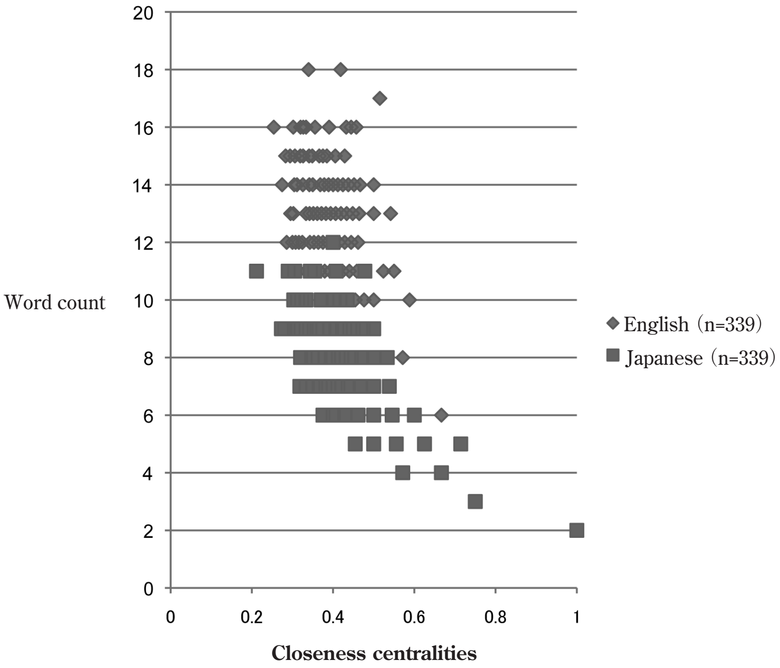

Figure 3.5 the distribution of closeness centralities and word counts (n = 339)

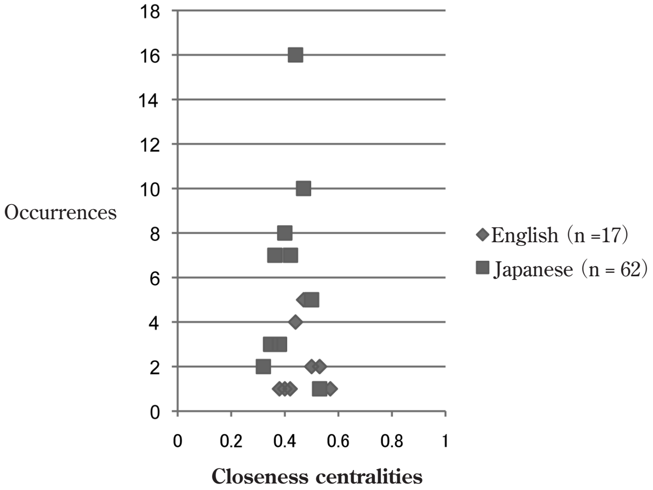

Figure 3.6 the closeness centralities and the occurrence of eight-word sentences in Basic300

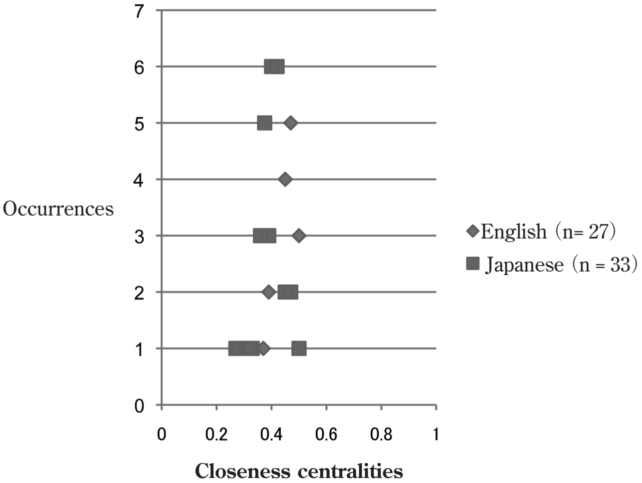

Figure 3.7 the closeness centralities and the occurrence of nine-word sentences in Basic300

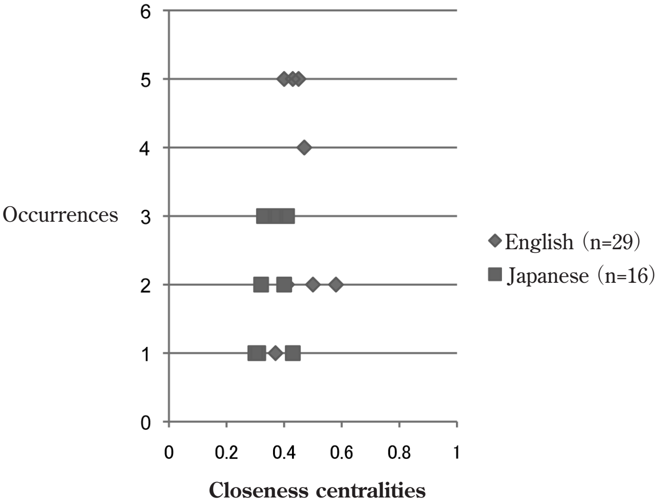

Figure 3.8 the closeness centralities and the occurrence of ten-word sentences in Basic300

3.4. Discussion

The different tendency between English sentences and Japanese ones in terms of the distribution of their degree centralities indicates that the typed-dependency trees for English sentences tend to have more varied structural settings than those for Japanese sentences in terms of their flatness. The distribution of their closeness centralities, on the other hand, does not show differences as obvious as that of their degree centralities. It also shows that closeness centralities are negatively correlated to the word count of sentences. These results can be interpreted that degree centralities reflect the structural differences between English and Japanese more explicitly than closeness centralities do, and degree centralities are less affected by the word count of each sentence than closeness centralities.

4. Conclusion

This study introduced the syntactic dependency structure of English and Japanese, and explored the possibility of graph-centrality measures to calculate the similarity between the syntactic dependency structure of an English sentence and that of Japanese sentence which semantically corresponds to the English sentence. The findings in this study on small-sized English-Japanese pairs should be examined using a larger-scale parallel corpus, which will be the topic of further study.

【Notes】

（1）The term “word” for Japanese means a syntactic unit（bunsetsu in Japanese language）, following Oya（2010a）. A syntactic unit of Japanese language consists of a content word with inflections or with particles.

（2）The names of dependency types for Japanese are based on Masuoka and Takubo（1992）.

（3）The term “word count” for Japanese means the number of syntactic units（Oya 2010a）.

【References】

De Marneffe, M.C., B. MacCartney and C. D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006. Retrieved on June 23, 2012 from http://nlp.stanford.edu/manning/papers/LREC_2.pdf

De Marneffe, M.C. and C. D. Manning. 2008. The Stanford typed dependencies representation. *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*. Retrieved on June 23, 2012, from http://nlp.stanford.edu/pubs/dependencies-coling08.pdf

De Marneffe, M.C. and C. D. Manning. 2011. *Stanford Typed Dependency Manual*. Retrieved June 21, 2012, from http://nlp.stanford.edu/software/dependencies_manual.pdf.

De Nooy, W., A. Mrvar and V. Batagelj 2005. *Exploratory Network Analysis with Pajek*. Cambridge University Press.

Debusmann, R. 2003. Dependency Grammar as Graph Description. : Prospects and Advances in the Syntax-Semantics Interface, Nancy. Retrieved July 27th, 2010, from http://www.ps.uni-saarland.de/~rade/papers/passi03.pdf

Debusmann, R. and M. Kuhlmann. 2007. *Dependency Grammar: Classification and Exploration. Project report* (*CHORUS, SFB 378*). Retrieved July 3, 2010, from http://www.ps.uni-saarland.de/~rade/papers/sfb.pdf

Freeman, L. 1979. Centrality in social networks. *Social Networks vol.1*, 215－239.

Hudson, R. 2010. *An Introduction to Word Grammar*. Cambridge University Press.

Iida, Y. 2010. *Eisakubun Kihon 300 Sen*. "300 Selected Basic Sentences for English Composition". Tokyo, Japan: Sundai

Kawahara, D. and S. Kurohashi. 2007. Daikibokakuframeni modozuku koubun kakukaisekino tougouteki kakuritumoderu "An integrated probabilistic model for syntactic and case analyses based on a large-scale case frames". Natural Language Processing vol. 14, no.4, 67－81.

Kurohashi, S. and M. Nagao. 1992. "A Method for Analyzing Conjunctive Structures in Japanese". Journal of Information Processing Society of Japan vol.33, no. 8. 1022－1031.

Kurohashi, S. and M. Nagao. 1994 A Syntactic Analysis Method of Long Japanese Sentences based on Coordinate Structure' Detection. Journal of Natural Language Processing vol.1, no.1. 35－57.

Kurohashi, S. and M. Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. Proceedings of the 1st International Conference on Language Resources and Evaluation. 719－724.

Masuoka, T. and Y. Takubo. 1992. *Kiso Nihongo Bumpo*. "Basic Japanese Grammar" Tokyo, Japan: Kuroshio Publishers.

Oya, M. 2009. A method of automatic acquisition of typed-dependency representation of Japanese syntactic structure. *Proceedings of the 14th Conference of Pan-Pacific Association of Applied Linguistics*. 337－340.

Oya, M. 2010a. Treebank-Based Automatic Acquisition of Wide Coverage, Deep Linguistic Resources for Japanese. M.Sc. thesis, School of Computing, Dublin City University.

Oya, M. 2010b. Directed acyclic graph representation of grammatical knowledge and its application for calculating sentence complexity. *Proceedings of the 15th International Conference of Pan-Pacific Association of Applied Linguistics*, 393－400.

Oya, M. 2012. Degree Centralities, Closeness Centralities, and Dependency Distances of Different Genres of Texts. *Proceedings of the 17th International Conference of Pan-Pacific Association of Applied Linguistics*. 89－90.

Tesnière, L. 1959. *Éléments de syntaxe structural*. Paris: Klincksieck.

Wasserman, S. and K. Faust. 1994. Social Network Analysis. Cambridge: Cambridge University Press.

（平成24年11月9日受理）