目白大学　人文学研究
第10号　2014年　205−215

Typed-dependency Tree Pairs of English and Japanese

大矢　政徳
Masanori OYA

Keywords : Typed-dependency trees, graph-centrality measures, syntactic similarity
*キーワード：*タイプ分けされた依存関係木、グラフ中心性指標、統語的類似性

1. Introduction

This study compares the syntactic dependency structures of English sentences and their Japanese counterparts in terms of their graph-centrality measures, which were proposed in Freeman (1979) and applied in syntactic typed dependency trees of English and Japanese by Oya (2010b, 2012, 2013), in order to explore the extent to which semantically similar sentences of the two languages share syntactic similarity. The structure of this study is as follows. Section 2 introduces the theoretical background: semantic similarity and syntactic similarity. Sections 3 and 4 report on the corpus-based experiment, and Section 5 concludes the study and provides suggestions for further study.

2. Semantic and syntactic similarity of English and Japanese sentences

It is assumed in this study that the structural properties of the syntactic dependency tree for an English sentence are not necessarily reflected in the structural properties of the syntactic dependency tree for the Japanese counterpart sentence.[1)] Consider example English sentence (1) and its Japanese counterpart (2) below.

(1)

"When you are told not to look at something, you become all the more eager to do so."

(2)

Miru-na-to	*iwa-reru-to*	*yokei-ni mitaku-naru-no-ga*	
look-NEG-POSTP	tell-passive-POSTP	more-POSTP	look.want-become-POSTP-POSTP

おおやまさのり：外国語学部英米語学科専任講師

ninjo-dearu.
human.nature-end

This sentence pair shows the largest difference between the word count of an English sentence and its Japanese counterpart in Basic 300 (Iida 2010). Figures 1 and 2 are the typed dependency trees for the counterpart sentences.[2)]

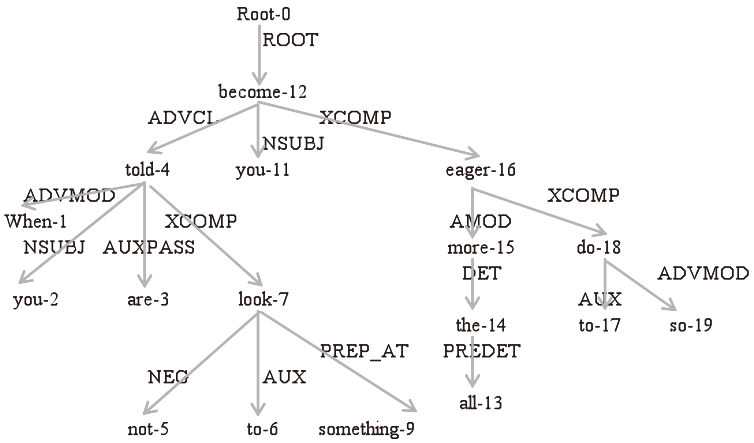

Figure 1. The typed dependency tree for "When you are told not to look at something, you become all the more eager to do so"

The word count of the English sentence is 18, and its degree centrality is 0.205.

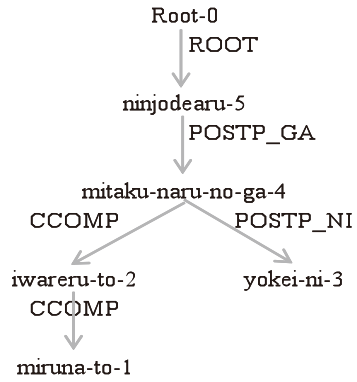

Figure 2. The typed dependency tree for "*Miruna-to iwareru-to yokei-ni mitaku-naru-no-ga ninjo-dearu.*"

The word count of the Japanese sentence is six, and its degree centrality is 0.4. The degree centrality of the syntactic typed dependency tree in Figure 1 is smaller than that in

Figure 2. Less flat syntactic typed dependency trees have smaller degree centralities[3]. Therefore, the syntactic typed dependency tree in Figure 1 is less flat than the one in Figure 2.

The insight gained from this comparison of the sentences in an English-Japanese translation pair is that the similarity of sentence meaning for an English sentence and its Japanese counterpart does not mean there are similar structural properties of their syntactic dependency trees. This insight seems to agree with our linguistic intuition, as the two languages are different and even belong to different language families.

However, this intuition does not help us to explain *how different* the sentences in each language are, in other words, in what ways their structural properties differ (e.g., Is a Japanese sentence *flatter* than its English counterpart, or *more embedded*?) and to what extent (e.g., *How flatter* is the Japanese sentence than its English counterpart?). With regard to these issues, graph centrality measures of typed dependency trees offer an explanation.

The difference between the word count of the English sentence and that of its Japanese counterpart is also taken into consideration, because it is possible that the structural difference between the two languages is subsumed into this aspect of syntactic property; that is, English may need more words than Japanese does in order to say something. If it is found that the differences between the degree centralities of the English sentences and those of the Japanese sentences in Basic 300 decrease in proportion to the differences between the word counts of the English sentences and those of the Japanese sentences (i.e., If it is found that, the more words an English sentence has than its Japanese counterpart sentence, the less flat the English syntactic typed dependency tree is in comparison to its Japanese counterpart), the word count can be considered to be the key factor for the difference of flatness (indicated by degree centrality). If, on the other hand, no relation is found between the difference in degree centralities and the difference in word counts of the sentences in Basic 300, the difference in word counts cannot be considered to be the key factor for the difference in flatness (indicated by degree centrality) of the English sentence and its Japanese counterpart. The same applies to the relation between word counts and embeddedness (indicated by closeness centralities).

3. Data analysis 1

3.1. Description of the data

The claim at the end of Section 2 must be verified, not by one English-Japanese translation pair above, but by a parallel corpus of English and Japanese. Some English-Japanese pairs in the parallel corpus are such that the English sentence and its Japanese

counterpart share relatively similar syntactic settings, while others may show structural diversity even though they share the same meaning. Therefore, this study uses Basic 300 (Iida 2010), which was also used in Oya (2013). Basic 300 contains 339 English-Japanese translation pairs. As the name indicates, the syntactic constructions of English sentences in the corpus are basic yet important ones, because they are compiled for Japanese high school students to memorize.

3.2. Procedure

The procedure in this study is essentially the same as that used in Oya (2013). First, the English sentences in Basic 300 are parsed by the 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. Next, the parse output for each sentence is manually corrected if any of the dependency relations or types in the parse output is found incorrect. 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 into the same format as that of Stanford Parser. The conversion policy is based on Oya (2010a). The parse output for each sentence is manually corrected if any of the dependency relations and types in the parse output is found incorrect. The centrality measures are calculated by the same Ruby script used for the English parse output.

Then, the following values are calculated for the English-Japanese counterpart sentences: difference in word count, difference in degree centrality, and difference in closeness centrality.

3.3. Results

Table 1 summarizes the average difference in word counts, degree centralities, and closeness centralities of the English-Japanese sentence pairs in Basic 300. The differences in degree centralities and closeness centralities are both less than zero, indicating that the degree centralities and closeness centralities of the English sentences in Basic 300 are smaller than those of their Japanese counterparts. This means that the English sentences in Basic 300 tend to be less flat and more embedded than their Japanese counterparts.

Table 1. The average differences in word counts, degree centralities, and closeness centralities between the English and Japanese sentences in Basic 300 (n = 339)

	w.c.	d.c.	c.c.
Basic 300	4.32	-0.07	-0.06

Note. w.c.: the average difference between the word counts of the English and Japanese sentences in Basic 300; d.c.: the average difference between the degree centralities of the English and Japanese sentences in Basic 300; c.c.: the average difference between the closeness centralities of the English and Japanese sentences in Basic 300.

For English-Japanese sentence pairs, the difference in degree centrality is not correlated with the difference in word count, as illustrated in Figure 3. Each dot in the plot represents an English-Japanese sentence pair. The x-axis represents the difference in word counts between the English sentence and its Japanese counterpart (the number of words in an English sentence minus the number of words in its Japanese counterpart sentence). The y-axis represents the difference in their degree centralities.

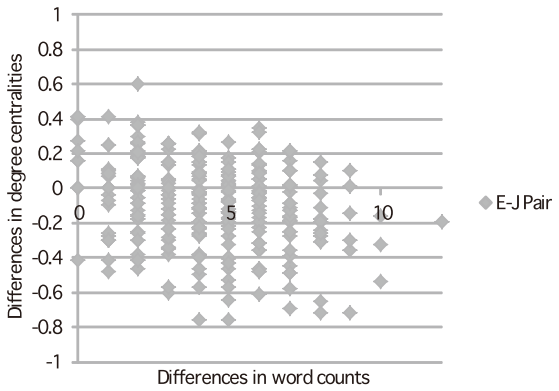

Figure 3. The differences in word counts and degree centralities of English sentences and their Japanese counterparts (n = 339)

Similarly, for English-Japanese sentence pairs, the difference in closeness centrality is not correlated with the difference in word count. In Figure 4, each dot and the x-axis represent the same information as in Figure 3, while the y-axis represents the difference in their closeness centralities.

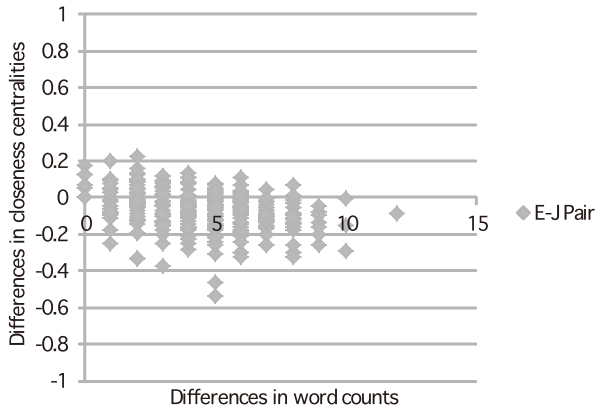

Figure 4. The differences in word counts and closeness centralities of English sentences and their Japanese counterparts (n = 339)

4. Data analysis 2

4.1. Background

In Section 3, we attempted to show that the typed dependency representation of English and Japanese and their graph-centrality measures can explain how different their syntactic structures are. As the results suggest, it seems plausible to argue that the difference in word counts is not the key factor in the difference of flatness (indicated by degree centrality) or difference of embeddedness (indicated by closeness centrality) of the English sentences and their Japanese counterparts. The next question here is which key factors contribute to the difference in flatness and embeddedness of English sentences and their Japanese counterparts.

In this study, we focus on the root word of the syntactic dependency tree. The assumption is that, if the roots of two syntactic trees are different, then their structural settings will also differ.

Here, we need to articulate the meaning of the word "different" according to the context of this study. In order to do this, this study articulates the idea of *lexical counterparts* at the root of the syntactic typed dependency tree. The Japanese lexical counterparts of an English word are Japanese words that can be the translation of the English word. For example, the Japanese lexical counterparts of the English word "eat" are *taberu, itadaku, meshiagaru, kurau*, etc.

It is assumed that, for an English-Japanese translation pair, if the root word of the English sentence is the lexical counterpart of the root word of the Japanese sentence, then the structural properties of their syntactic typed dependency trees are similar, if not

identical. If, on the other hand, the root word of its English sentence is not the lexical counterpart of the root word of the Japanese sentence, then the structural properties of their syntactic typed dependency trees are different.

4.2. Procedure

We verify the assumption introduced in Section 4.1 as follows. First, the English-Japanese translation pairs in Basic 300 are divided into the following two groups: (1) the sentence pairs in which the root word of the English sentence is the lexical counterpart of the root word of the Japanese sentence (henceforth *sameRoot*) and (2) the sentence pairs in which the root word of the English sentence is not the lexical counterpart of the root word of the Japanese sentence (henceforth *differentRoot*).

Then, the following figures are compared across the two groups: the average of the difference in word counts and the average of the difference in degree centralities. If the average of the difference in word counts of the English sentences and their Japanese counterparts in sameRoot is significantly smaller than that in differentRoot, then the choice of the root word can be considered to result in the difference of their word counts. If the average of the difference in degree centralities between the English sentences and their Japanese counterparts in sameRoot is significantly smaller than that in differentRoot, then the choice of the root word can also be considered to result in the difference of their degree centralities. The same will apply for their closeness centralities.

4.3. Results

The sameRoot group contains 115 English-Japanese sentence pairs (approximately 34% of Basic 300), while the differentRoot group contains 224 pairs (approximately 66%). This simple fact suggests that English and Japanese tend not to share the same root ("to share the same root" here means that the root of the English sentence is a lexical counterpart of the root word of the Japanese sentence).

Table 2 summarizes the results of the comparisons between sameRoot and different Root. There is no obvious difference between the two groups.

Table 2. The average differences in word counts, degree centralities, and closeness centralities of the English and Japanese sentences in sameRoot and differentRoot

	w.c.	d.c.	c.c.
sameRoot	4.23	-0.09	-0.08
differentRoot	4.37	-0.07	-0.06

Note. w.c.: the average difference between the word counts of the English and Japanese sentences in each group; d.c.: the average difference between the degree centralities of the English and Japanese sentences in each group; c.c.: the average difference between the closeness centralities of the English and Japanese sentences in each group.

Figure 5 plots the differences in word counts and degree centralities of the sentence pairs in sameRoot.

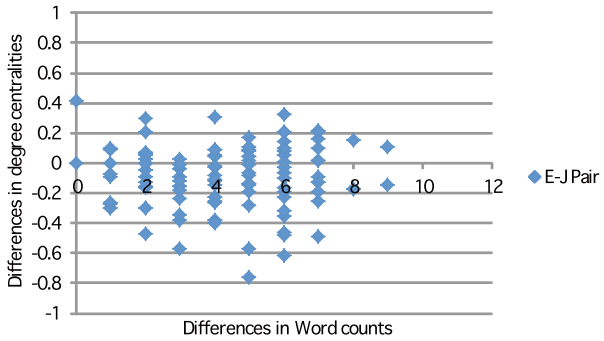

Figure 5. The differences in word counts and degree centralities of English sentences and their Japanese counterparts whose roots are lexical counterparts (n = 115)

Figure 6 shows the differences in word counts and degree centralities in differentRoot.

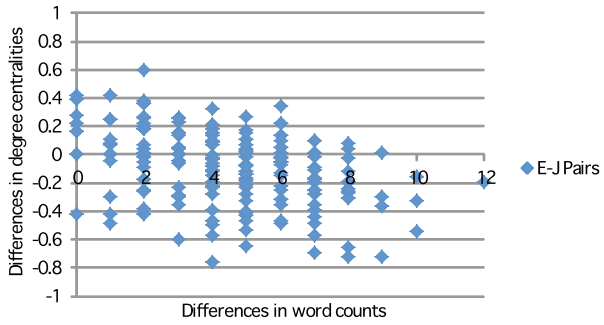

Figure 6. The differences in word counts and degree centralities of English sentences and their Japanese counterparts whose roots are NOT lexical counterparts (n = 224)

As Figures 5 and 6 illustrate, whether the root word of the English sentence is a lexical counterpart of the root word of the Japanese sentence does not seem to be reflected in the distribution of the difference in word counts or the distribution of the difference in degree

centralities.

Figure 7 plots the differences in word counts and closeness centralities of the sentence pairs in sameRoot.

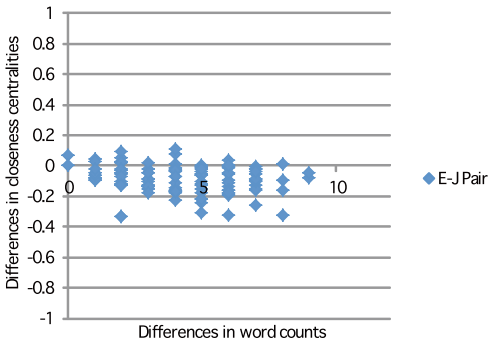

Figure 7. The differences in word counts and closeness centralities of English sentences and their Japanese counterparts whose roots are lexical counterparts (n = 115)

Figure 8 shows the differences in word counts and closeness centralities in different Root.

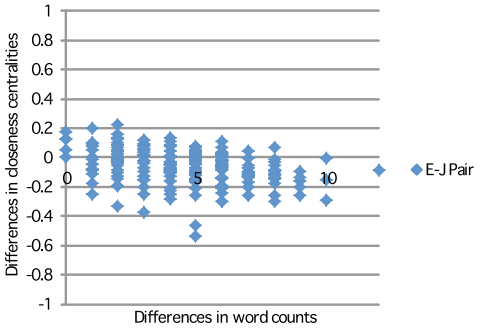

Figure 8. The differences in word counts and closeness centralities of English sentences and their Japanese counterparts whose roots are NOT lexical counterparts (n = 224)

As Figures 7 and 8 illustrate, whether the root word of the English sentence is a lexical counterpart of the root word of the Japanese sentence does not seem to be reflected in the distribution of the difference in word counts or distribution of the difference in closeness centralities.

5. Conclusion

This study compared the syntactic dependency structures of English sentences and their Japanese counterparts in terms of their graph-centrality measures, which were proposed in Freeman (1979) and applied in syntactic typed dependency trees of English and Japanese by Oya (2010b, 2012, 2013), in order to explore the extent to which semantically similar sentences of the two languages share syntactic similarity. The data analyses indicate that (1) English sentences in the parallel corpus tend to have less flat and more embedded structural settings, (2) the difference in word counts between the English sentences and their Japanese counterparts is not related to the difference of the structural settings of the syntactic typed dependency tree in terms of their flatness (indicated by degree centrality) or embeddedness (indicated by closeness centrality), and (3) the difference of the root word is not related to the difference of the structural settings, which is the same as (2) above. This study can be extended by (1) considering more than one translation counterpart Japanese sentence for each English sentence, (2) dividing the data of Basic 300 in terms of factors other than the same/different root, and (3) taking Japanese elliptic sentences into consideration. Each of these will be a topic of further study.

Notes

1) This study does not deny that one English sentence can be translated into more than one Japanese sentence. However, it concentrates on English-Japanese sentence pairs in the small-scale parallel corpus used in Oya (2013). It will be interesting to compare one English sentence with more than one Japanese translation of it in terms of the centrality measures, which will be the focus of future research.
2) In this study, periods and commas are not included in the syntactic dependency trees of English or Japanese.
3) For the definition of degree centrality, see Freeman (1979) and Wassermann & Faust (2004). For the meaning of degree centrality and closeness centrality in terms of syntactic typed dependency trees, see Oya (2010b, 2012, 2013).

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.
Freeman, L. 1979. Centrality in social networks. *Social Networks* vol. 1, 215−239.
Kawahara, D. and S. Kurohashi. 2007. Daikibokakuframeni modozuku koubun kakukaisekino tougouteki kakuritumoderu "An integrated probabilistic model for syntatic 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 Stuctures 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.
Iida, Y. 2010. *Eisakubun Kihon 300 Sen* [300 Selected Basic Sentences for English Composition]. Tokyo: Sundai.
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.
Oya, M. 2013. Syntactic dependency structures of English and Japanese. *Mejiro Journal of Humanities* vol. 9, 151−164.
Wasserman, S. and K. Faust. 1994. Social Network Analysis. Cambridge: Cambridge University Press.

（平成25年11月6日受理）