# Degree Centralities, Closeness Centralities, and Dependency Distances of Different Genres of Texts

Masanori Oya

Mejiro University
m.oya@mejiro.ac.jp

## Abstract
This study investigates whether the directed acyclic graph representations of the typed-dependency trees for the sentences (typed-dependency DAGs) in different genres of texts in the manually annotated sub corpus of American National Corpus (MASC 500k) show different distribution of their degree centralities, closeness centralities and dependency distances. Different distributions of degree centralities and closeness centralities are found among different genres, yet this can be considered to be a result of the fact that texts with shorter sentences tend to have larger degree centralities and closeness centralities.

## Keywords
Dependency grammar, Stanford Dependency Parser, Graph centralities

## 1 Introduction
Oya (2010b) showed that small-scale corpora of different genres of texts have different distributions of degree centralities and closeness centralities of the directed acyclic graph of the typed-dependency trees for the sentences (typed-dependency DAGs: Oya2010a), and Oya (2011) showed that the corpora of the similar scale have different distributions of dependency distances. This study uses a corpus larger than these corpora used in Oya (2010b) and Oya (2011) in terms of the number of sentences, and wider in terms of genres, in order to examine whether large-scale corpora of different genres of texts have different distributions of degree centralities, closeness centralities and dependency distances of the typed-dependency DAGs. Section 2 explains the background of this study; dependency grammar, graph centralities and dependency distance. Section 3 deals with the analysis of data used in this study, and with the discussion of the results, and Section 4 concludes this study.

## 2 Background
### 2.1 Dependency grammar
Lucien Tesnière, a French linguist in the 20th century, is considered to be the father of dependency grammar. He argues that sentence structures are defined by the relationships between the head word and tail word; each word in a sentence is dependent on another word, no word in a sentence is independent, and the dependency relationship between words is characterized by a governor and a dependent (Tesnière 1959).

There are several different dependency-grammar frameworks based on Tesnière's idea: Link Grammar (Sleator and Temperley 1991), Extensible Dependency Grammar (Debusmann 2003; Debusmann and Kuhlmann 2007), Word Grammar (Hudson 2010), and Stanford Dependencies (de Marneffe and Manning 2008).

The Dependency-grammar framework used in this study is Stanford Dependencies, because it is the linguistic basis of Stanford Parser (de Marneffe and Manning 2006, 2008, and 2012) which I also use to parse a large number of sentences in the corpus in this study. Stanford Dependencies contain 55 dependency types (de Marneffe and Manning 2012). For example, the figure below is the typed-dependency tree for a sentence "Sarah has read this book."

In this typed-dependency tree, the word "Sarah" depends on the verb "read", and this dependency is typed as "NSUBJ", which means "nominal subject". The word "has" depends on "read" with the type "AUX" (auxiliary), the word "book" on "read" with the type "DOBJ" (direct object), and the word "this" on "book" with "DET" (determiner). The main verb "read" depends on Root. The presence of Root in a typed-dependency tree ensures that all the words in a sentence depend on another entity, and also reflects the dependency of the main predicate ("read" in the tree above) to a sentence-external, discourse level of representation.

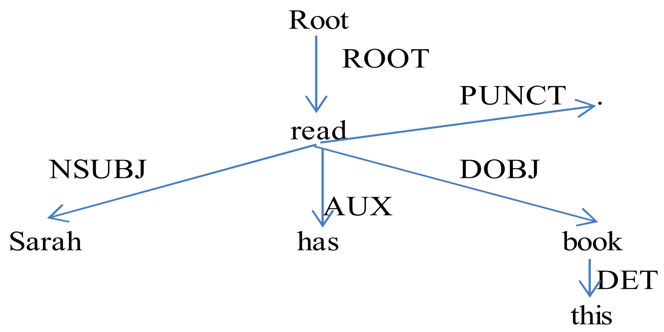


*Figure 1*: The typed-dependency tree for “Sarah has read this book.”

### 2.2 Typed-dependency trees as graphs

Oya (2010a, 2010b, and 2011) assume that the typed-dependency tree for a sentence can be considered to be a graph in the sense of graph theory which is a part of mathematics (Freeman, 1978; Wasserman & Faust, 1994). A graph consists of a set of *vertices* and a set of *arcs* connecting these vertices. An arc is *directed* if it starts from one vertex to another. The number of arcs attached to a vertex is the *degree* of the vertex. In a typed-dependency tree, each word in the tree is a vertex, and the dependency relation between words is an arc which starts from the head to the tail. Each dependency relation is labeled with a *dependency type* which indicates the function that the tail has with respect to the head of the dependency relation.

### 2.3 Graph centrality

In graph theory, various indices have been defined to show the structural property of a given graph. These indices can be used to show the structural property of a given typed-dependency tree (Oya 2010b and 2011). They allow us to understand the structural property of a given sentence more objectively. Among these indices, Oya (2010b) proposed to use *graph centrality* for an index to show the complexity of typed-dependency trees. Graph centrality shows the relative importance of vertices in a given graph (Freeman, 1978; Wasserman & Faust, 1994).

#### 2.3.1 Degree centrality

*Degree centrality* indicates the extent to which the vertices in a graph are concentrated to one particular vertex. It is calculated by the following formula (Wasserman & Faust 1994: 180). The degree centrality of a given graph ($C_D$ in the formula below) is the sum of the maximum degree in the graph minus the degree of each of all the other vertices, divided by the largest possible sum of the maximum degree of the graph minus the degree of all the other vertices. In the formula below, $g$ is the number of vertices in a graph, $C_D(n^*)$ is the largest degree among the vertices in the graph, and $C_D(n_i)$ is the degree of the $i$th vertex of the graph.

$$C_D = \frac{\sum_{i=1}^{g}[C_D((n^*)-C_D(n_i))]}{\max\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]} \qquad (2.1)$$

The denominator equals the number of vertices minus 2 multiplied by the number of vertices minus one (See Freeman 1978).

The degree centrality of a given typed-dependency tree indicates how flat the tree is (Oya 2010b). The flatness of a tree means the extent to which one particular word is the dependency head of other words. Degree centrality increases in proportion to the flatness of typed-dependency trees. For example, consider the example sentence “Sarah has read this book.” This sentence contains seven words, including Root and the period. The maximum degree in the typed-dependency graph for this sentence is 5 at the word “read”; the sum of the maximum degree in the graph minus the degree of each of all the other vertices is (5-1)+(5-1)+(5-5)+(5-1)+(5-1)+(5-2)+(5-1)=23. The denominator is 30; hence the degree centrality of the typed-dependency tree for the sentence “Sarah has read this book.” is 23/30≒0.767.

Next, consider a sentence “Sarah would have read this book.” Figure 2 is the typed-dependency tree for this sentence.

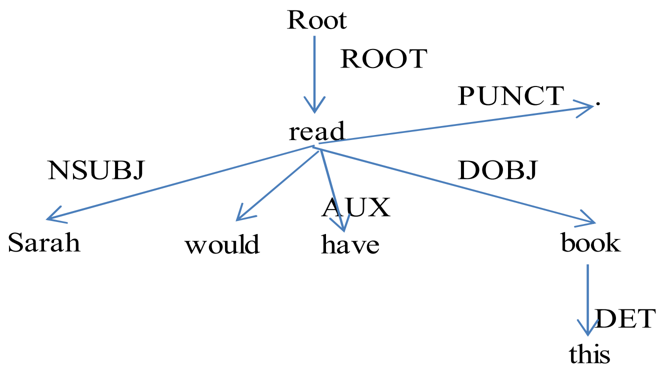


*Figure 2*: The typed-dependency tree for "Sarah would have read this book."

This sentence contains eight words, including the Root and the period. The maximum degree in the typed-dependency tree for this sentence is 6 at the word "read"; the sum of the maximum degree in the graph minus the degree of each of all the other vertices is (6-1)+(6-1)+(6-1)+(6-6)+(6-1)+(6-1)+(6-2)+(6-1)=34. The denominator is 42; hence the degree centrality of the typed-dependency tree for the sentence "Sarah would have read this book." is 34/42≒0.81. As the degree centralities of these examples sentences show, a typed-dependency tree with a flatter setting has a larger degree centrality.

Degree centralities of typed-dependency trees can indicate the flatness of sentences across languages. Consider a typed-dependency tree for an English sentence in Figure 3 and its Japanese equivalent in Figure 4.

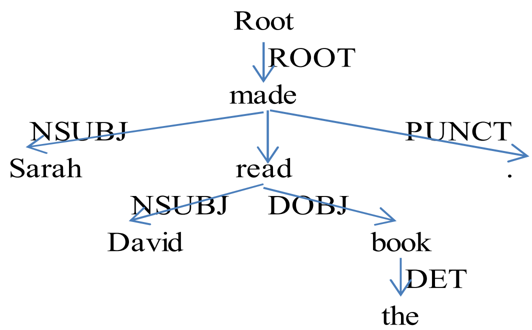


*Figure 3*: The typed-dependency tree for "Sarah made David read the book."

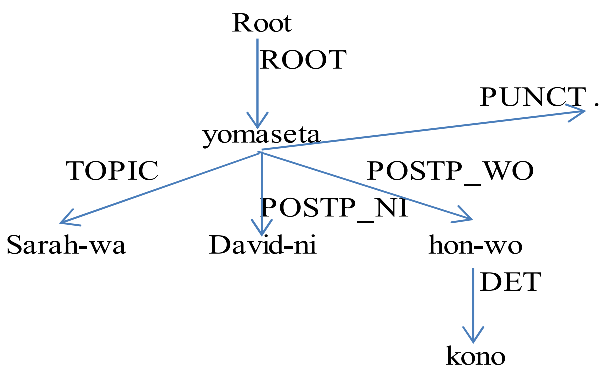


*Figure 4*: The typed-dependency tree for "Sarah-wa David-ni kono hon-wo yomaseta (Sarah made David read this book.)"

The degree centrality of the typed-dependency tree in Figure 3 is about 0.428, while that in Figure 4 is about 0.766. As this illustration suggests, the larger degree centrality of the typed-dependency tree in Figure 4 numerically indicates that this tree is flatter than its English counterpart.

### 2.3.2 Closeness centrality

*Closeness centrality* is defined as the reciprocal of the sum of the length of a path from one vertex to another in a graph (Freeman 1978; Wasserman and Faust 1994)[1]. This calculation is represented in the following formula (Sabidussi 1966; Wasserman and Faust 1994: 184), in which $g$ means the number of vertices (or *nodes*), and $d(n_i, n_j)$ is the shortest path (*geodesic distance*) between the vertex $n_i$ and vertex $n_j$.

$$C_c = (n_i)\frac{1}{\sum_{j=1}^{g} d(n_i, n_j)} \tag{2.2}$$

Wasserman and Faust (1994: 185) point out that the maximum value attained by the formula (2.2) above depends on the number of vertices in a graph, and therefore it is difficult to compare values across networks of different sizes. Therefore, they refer to Beauchamp (1965) which suggests to use standardized indices calculated by the following formula.

$$C_c = (n_i)\frac{g-1}{\sum_{j=1}^{g} d(n_i, n_j)} \tag{2.3}$$

Wasserman and Faust (1994) point out that this can be viewed as the inverse average distance between vertex $i$ and all the other vertices, and it ranges from 0 to 1; it equals 1 when a vertex is *adjacent* (connected by one edge) to all the other vertices.

In a typed-dependency tree representation for a sentence, the relevant length is that between the root and all the other words in the tree, because it represents the depth of embedding of each word from the root in the sentence (Oya 2010b).

Closeness centrality decreases in proportion to the embeddedness of typed-dependency trees. For example, the example sentence "Sarah has read this book." has six paths from the Root; Root-read, Root-read-Sarah, Root-read-has, Root-read-., Root-read-book, and Root-read-book-this. The lengths of these paths are 1, 2, 2, 2, 2, and 3, respectively (the starting vertex is not included). The average length of them is 2, and the closeness centrality of this sentence is the inverse of 2, that is, 0.5.

Next, consider an example sentence "My brother has read this book." Figure 5 is the typed-dependency tree for this sentence.

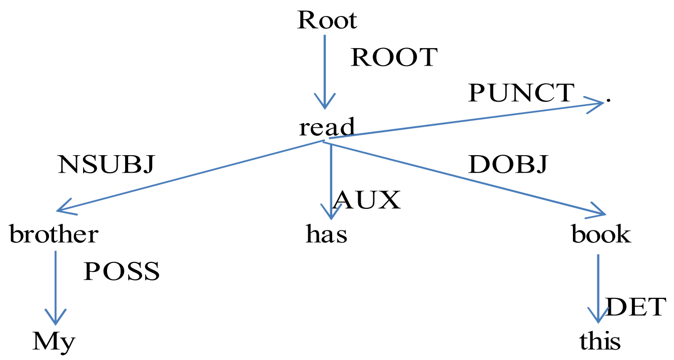


*Figure 5*: the typed-dependency tree for "My brother has read this book."

This sentence has seven paths from the Root; Root-read, Root-read-brother, Root-read-brother-My, Root-read-has, Root-read-., Root-read-book, and Root-read-book-this. The lengths of these paths are 1, 2, 3, 2, 2, 2, and 3. The average length of them is 15/7≒2.142, whose inverse is the closeness centrality of this sentence; 1/2.142≒0.467.

Lastly, consider an example sentence "Sarah read the books David has." Figure 6 is the typed-dependency tree for this sentence.

[1] Oya (2010b) proposed the term *path length* for the average length of a path from one given node to another in a graph. However, unlike degree centrality, path lengths do not fall between 0 to 1. Therefore, Oya (2011) used the concept of closeness centrality as defined by Beauchamp (1965).

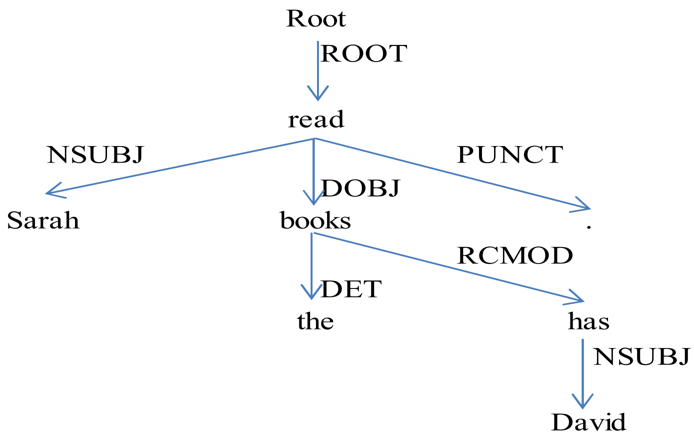


*Figure 6*: The typed-dependency tree for "Sarah read the books David has."

This sentence has seven paths from the Root; Root-read, Root-read-Sarah, Root-read-books-the, Root-read-books, Root-read-books-has-David, Root-read-books-has, and Root-read-.. The lengths of these paths are 1, 2, 3, 2, 4, 3, and 2. The average length of them is 17/7≒2.44, whose inverse is the closeness centrality of this sentence; 1/2.44≒0.41. As the closeness centralities of these example sentences show, a typed-dependency tree with a more embedded setting has a smaller closeness centrality.

Oya (2010b) showed that the distributions of degree centralities and that of closeness centralities of the sentences in different small-sized corpora (English essays written by Japanese university students, abstracts of academic journals, and the 1st chapter of "*The Golden Bough*") are different: the English essays tend to contain sentences with larger degree centralities than the abstracts; the abstracts tend to contain sentences with smaller closeness centralities than the English essay.

### 2.4 Dependency distance

The drawback of Graph-centrality measures for sentence complexity is that they do not show the *dependency distance* between the head and the tail of a dependency relation because they abstract away the linear order of words in a sentence (Oya 2011). Dependency distance is defined as the number of words between the head and the tail of a dependency relation, and the average dependency distance of a sentence is the sum of all the dependency distance divided by the number of dependency relations. For example, in the typed-dependency tree in Figure 1, the dependency distance between the word "Sarah" and "read" is 2, "has" and "read" is 1, "book" and "read" is 2, "this" and "book" 1, "read" and "." is 4. The Root is considered as the 0th word; hence the dependency distance between "Root" and "read" is 3. The average dependency distance of the sentence "Sarah has read this book." is (2+1+2+1+4+3) / 6≒2.167.

Dependency distance of a given typed-dependency tree can be employed to indicate the complexity of the tree. With respect to dependency distance, Gibson (1998, 2000) argued that the syntactic complexity of sentences increases in proportion to the dependency distance, and Temperley (2006) proposed the presence of preference for longer or shorter dependency distances according to different syntactic contexts. Oya (2011) argued that the average dependency distance of the sentences can be applied to calculate the sentence complexity of the English sentences written by native speakers of English and that of those written by non-native speakers of English, and conducted an analysis of average dependency distance of sentences in small-sized corpora taken from different types of writers (English textbooks for Japanese high schools, English essays written by Japanese university students, and newspaper articles). The results showed that the average dependency distance of the sentences in newspaper articles is the largest among these three corpora.

### 2.5 Related work

Since the seminal work by Biber (1988), there have been a number of attempts to use syntactic features for genre classification. For example, part-of-speech trigrams are used by Argamon et al. (1998), and Santini (2004) applied their method to ten different genres in BNC. Their method uses only shallow syntactic information (POS tags), and it is expected that deep syntactic information will also be effective for genre classification. Deep syntactic information includes the dependency structure of a sentence in terms of degree centrality and closeness centrality, and the average dependency distances of a sentence.

# 3 Data analysis

## 3.1 Data description

The corpus used in this study is the manually annotated sub corpus of American National Corpus (MASC 500k). This corpus contains approximately 500,000 words of contemporary American English, drawn from Open American National Corpus (OANC) (Ide and Suderman 2004). MASC 500k covers a wide range of genres: blogs, essays, fictions, short fictions taken from a website Ficlet (now closed), government documents, jokes, journals, newspapers, non-fictions, technical reports, and travel guides. Texts of emails, spam emails, movie scripts, speeches, and debates are also included, but not included in this study. Table 1 shows the descriptive statistics of the sentence numbers, word counts and words per sentence (WPS) of each genre.

Table 1: The total number of sentences, the total number of words and the mean length of a sentence in each genre

| Subsections | Sentences | Words | WPS Mean | WPS S.D. |
|---|---|---|---|---|
| Blog | 1524 | 28381 | 18.62 | 12.34 |
| Essay | 1072 | 27367 | 25.52 | 13.18 |
| Ficlets | 2645 | 30555 | 13.34 | 7.15 |
| Fiction | 2639 | 37531 | 14.22 | 8.32 |
| Govt-doc | 1028 | 24277 | 23.61 | 12.55 |
| Jokes | 2254 | 31751 | 14.08 | 8.6 |
| Journal | 867 | 21997 | 25.37 | 14.45 |
| News | 1196 | 26877 | 22.47 | 10.43 |
| Non-Fiction | 1278 | 26441 | 20.68 | 11.41 |
| Technical | 825 | 19787 | 23.98 | 13.58 |
| TravelGuide | 1196 | 24187 | 20.23 | 8.6 |
| Total | 16524 | 299151 | | |

The subsections Fiction, Ficlets and Jokes show relatively smaller mean WPSs compared to other subsections. Their standard deviations are also smaller than those of other sections. The subsection Essay has the largest WPS among them.

## 3.2 Method and Results

The raw texts in each of the genres without tags (downloaded as a data-only file from the website of ANC: http://www.anc.org/MASC/Download.html) are parsed by Stanford Parser, and the degree centrality, closeness centrality and dependency distance of the parse-output typed-dependency DAGs for the sentences in the texts are calculated automatically by scripts written in Ruby. Table 2 shows the descriptive statistics of degree centralities, closeness centralities and average dependency distance of each genre. The results show that different genres of texts show different distribution of degree centrality, closeness centrality and average dependency distance.

Table 2: Descriptive statistics of degree centralities, closeness centralities, and dependency distance

| | Degree | | Closeness | | Dep.Dist. | |
|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Blog | 0.43 | 0.26 | 0.39 | 0.12 | 2.99 | 1.19 |
| Essay | 0.26 | 0.19 | 0.32 | 0.09 | 3.58 | 1.09 |
| Ficlets | 0.57 | 0.29 | 0.47 | 0.11 | 2.31 | 1.01 |
| Fiction | 0.54 | 0.27 | 0.43 | 0.11 | 2.65 | 1.05 |
| Govt-doc | 0.26 | 0.18 | 0.33 | 0.10 | 3.41 | 1.12 |
| Jokes | 0.51 | 0.27 | 0.44 | 0.13 | 2.55 | 1.15 |
| Journal | 0.27 | 0.19 | 0.33 | 0.09 | 3.63 | 1.35 |
| News | 0.27 | 0.17 | 0.33 | 0.09 | 3.34 | 0.94 |
| Non-Fiction | 0.37 | 0.22 | 0.36 | 0.10 | 3.20 | 1.07 |
| Technical | 0.32 | 0.22 | 0.34 | 0.12 | 3.37 | 1.27 |

The subsections Ficlets, Fiction and Jokes are the three subsections with the top-three largest mean degree centrality. Their standard deviations are also larger than those of other sections. The subsection Essay shows the smallest degree centrality among them.

The subsections Ficlets, Fiction and Jokes are the three subsections with the top-three largest mean closeness centrality. The standard deviation of Jokes is the largest of them, but those of Blog and Technical are larger than those of Ficlets and Fiction.

The subsections Fiction, Ficlets and Jokes are the three subsections with the top-three shortest dependency distance. The subsection Journal has the longest dependency distance. The subsection Journal has the largest standard deviation, and News the smallest.

Example typed-dependency trees taken from these subsections will illustrate the claim that flatter trees have larger degree centralities and more embedded trees have smaller closeness centralities. For example, Figure 7 below is the typed-dependency tree for an example sentence selected from the subsection Journal. This has 10 words. Its degree centrality is about 0.722, its closeness centrality is about 0.473, and its average dependency distance is about 2.444.

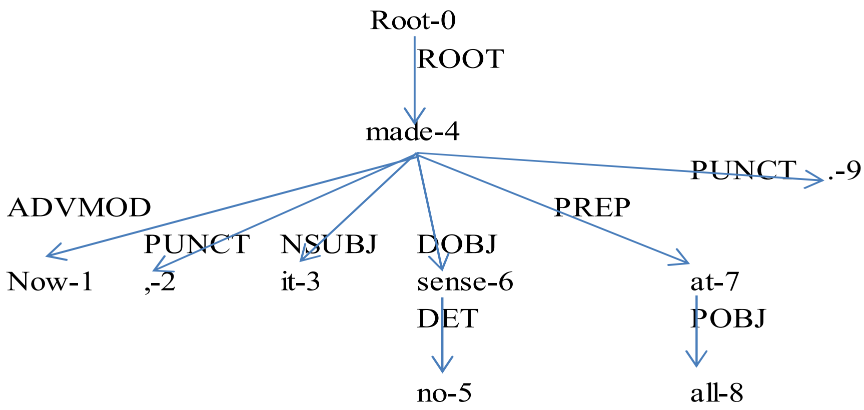


*Figure 7*: The typed-dependency tree for “Now, it made no sense at all.”

Figure 8 is the typed-dependency tree for another example sentence selected from the subsection Journal. This has 10 words. Its degree centrality is about 0.444, its closeness centrality is 0.391, and its average dependency distance is about 2.333.

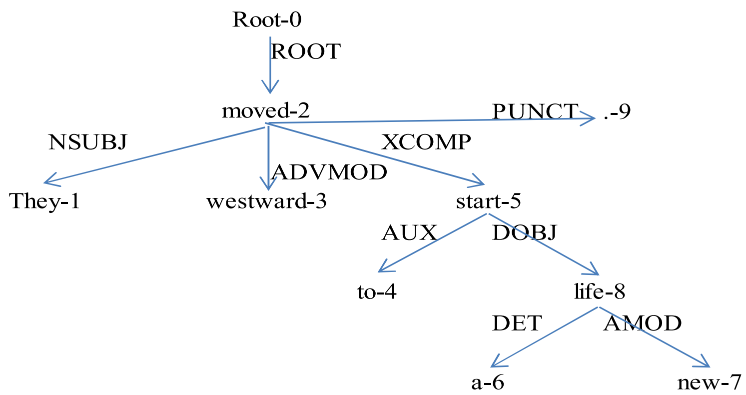


*Figure 8*: The typed-dependency tree for “They moved westward to start a new life.”

Figure 9 is the typed-dependency tree for an example sentence selected from the subsection Blog. This has 10 words. Its degree centrality is about 0.722, its closeness centrality is 0.45, and its average dependency distance is about 2.666.

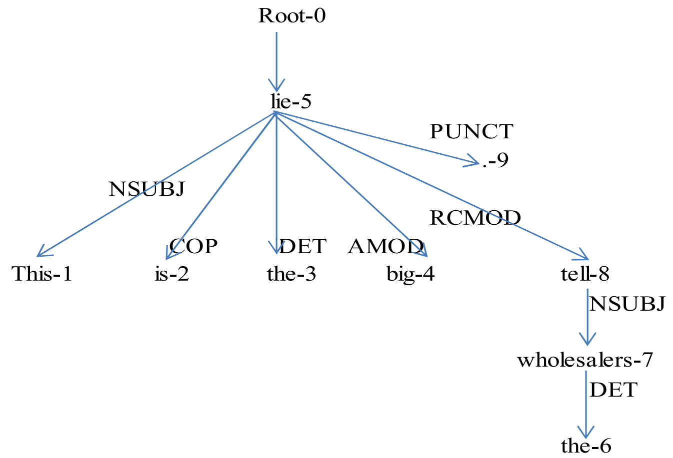


*Figure 9*: The typed-dependency tree for "This is the big lie the wholesalers tell."

Figure 10 is the typed-dependency tree for another example sentence selected from the subsection Blog. This has 10 words. Its degree centrality is about 0.305, its closeness centrality is 0.36, and its average dependency distance is about 2.555.

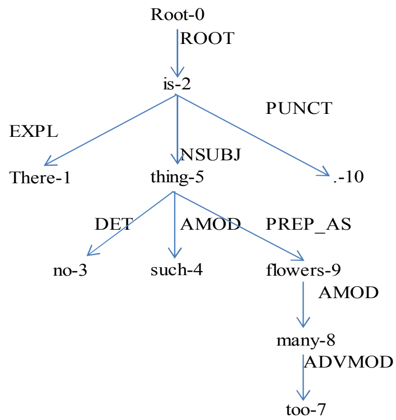


*Figure 10*: The typed-dependency tree for "There is no such thing as too many flowers."

The degree centralities, closeness centralities, and average dependency distances of these example typed-dependency trees are shown in Table 3 below. The typed-dependency tree in Figure 7 is as flat as that in Figure 9, because they have the same degree centrality. The typed-dependency tree in Figure 7 is the least embedded than others, because it has the largest closeness centrality among them. The typed-dependency tree in Figure 10 is the least flat and the most embedded one among them, because it has the smallest degree centrality and the smallest closeness centrality. The average dependency distances of these typed-dependency trees do not seem to be as different as their degree centralities and their closeness centralities.

Table 3: The degree centralities, closeness centralities, and average dependency distances of the example typed-dependency trees

| | Degree | Closeness | Dep.Dist |
|---|---|---|---|
| Figure 7 | 0.722 | 0.473 | 2.444 |
| Figure 8 | 0.444 | 0.391 | 2.333 |
| Figure 9 | 0.722 | 0.450 | 2.666 |
| Figure 10 | 0.305 | 0.360 | 2.555 |

The larger degree centrality and closeness centrality of the sentences in Fiction, Ficlet and Jokes can be the result of the fact that they contain sentences shorter than those in other genres on average, as their WPSs indicate; degree centrality tends to become smaller in proportion to the number of words in sentences (Satoshi Yoshida, p.c.). This suggests that the difference of distributions of degree centrality and closeness centrality among sentences in different genres is nothing but a paraphrase of different distribution of WPSs among these sentences. In order to address this issue, it will be desirable to treat degree and closeness centrality with precaution; for example, controlling the number of words in a sentence to examine how both centralities of the sentences of equal word count show different distributions. Figure 11 is the distribution of degree centralities of 10-word sentences in each genre of MASC500k, and Figure 12 is the distribution of degree centralities of 20-word sentences in the same corpus.

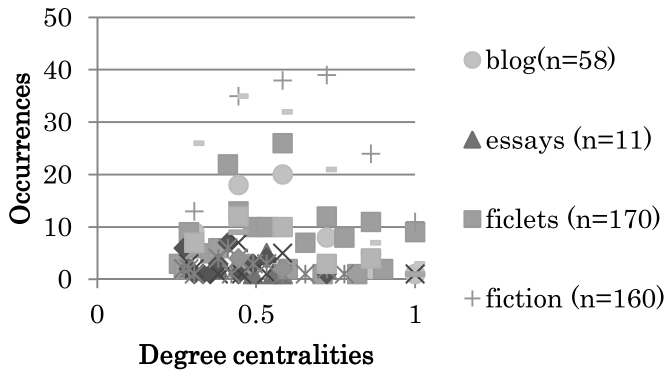


*Figure 11*: the distribution of degree centralities of 10-word sentences in each genre of MASC500k

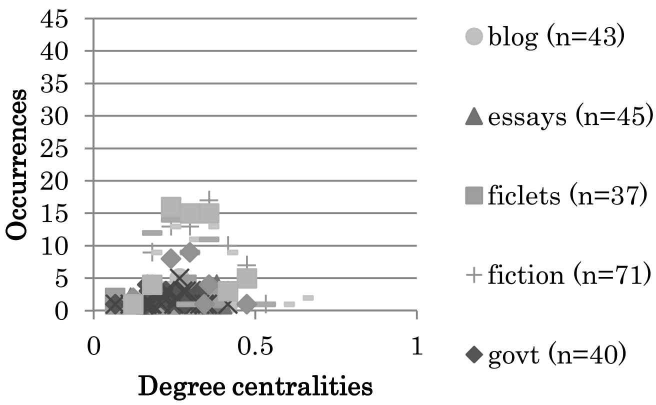


*Figure 12*: the distribution of degree centralities of 20-word sentences in each genre of MASC500k

Notice that the distribution of degree centralities of 10-word sentences are more dispersed than that of 20-word sentences, and in the distribution of 20-word sentences, there is only a small number of degree centralities more than 0.5.

Sentences in different genres show different distributions of degree centralities. We can see the difference more explicitly if we look at one particular degree centrality across different genres. For example, let us concentrate on Fiction and Journal. 39 sentences of all the 10-word sentences in Fiction (n=160) have the degree centrality 0.72. This means that about 24% of these sentences in Fiction have the degree centrality 0.72.

On the other hand, 2 sentences of all the 10-word sentences in Journal (n=28) have the degree centrality 0.72. This means that about 7% of 10-word sentences in Journal have the degree centrality 0.72. These results suggest that sentences in Fiction tend to be flatter than those in Journal, as far as 10-word sentences in these genres are concerned.

17 sentences of all the 20-word sentences in Fiction (n=71) have the degree centrality 0.35. This means that about 24% of these sentences in Fiction have the degree centrality 0.35. On the other hand, 1 sentence of all the 20-word sentences in Journal (n=35) has the degree centrality 0.35. This means that about 2% of these sentences in Journal have the degree centrality 0.35. Again, these results suggest that sentence in Fiction tend to be flatter than those in Journal, as far as 20-word sentences are concerned.

The distributions of closeness centralities of the sentences of the same word count are somewhat different from those of degree centralities. Figure 13 is the distribution of closeness centralities of 10-word sentences, and Figure 14 is the distribution of closeness centralities of 20-word sentences.

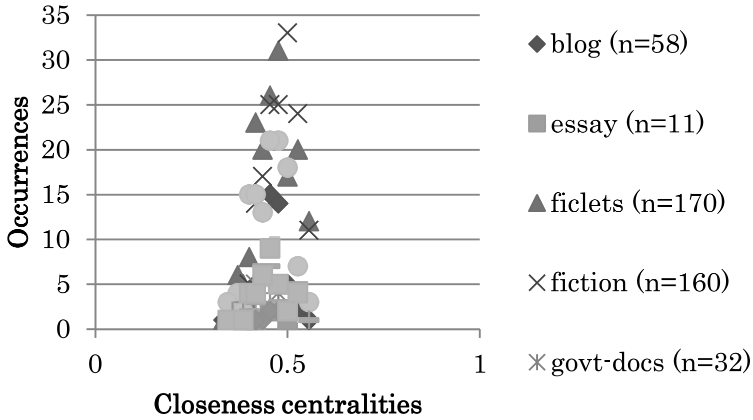


*Figure 13*: the distribution of closeness centralities of 10-word sentences in each genre of MASC500k

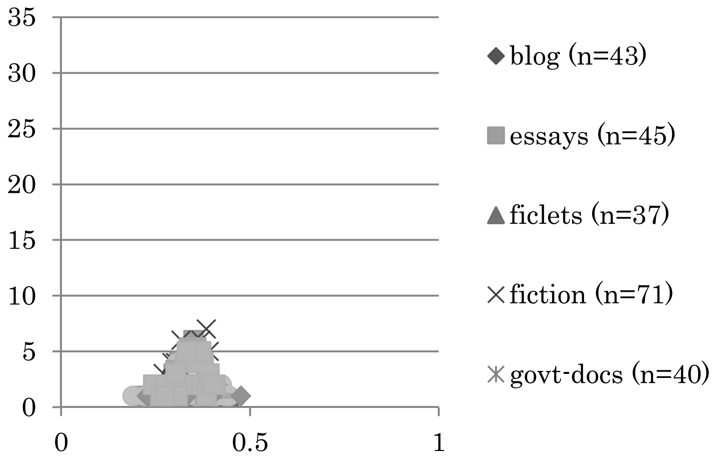


*Figure 14*: the distribution of closeness centralities of 20-word sentences in each genre of MASC500k

Notice that the distribution of closeness centralities of 10-word sentences are more dispersed than that of 20-word sentences, and there is no closeness centrality more than 0.5 in the distribution of 20-word sentences. On the other hand, the number of different closeness centralities increases in proportion to the increase in the word counts; for example, as for the 10-word sentences in Fiction, there are 11 different values of closeness centralities, while for the 20-word sentences in Fiction there are 26 different values of closeness centralities. The closeness centralities of 10-word sentences and those of 20-word sentences in Fiction are shown in Tables 4 and 5 below.

Table 4: the different values of closeness centrality and the number of sentences which have the same closeness centrality value (10-word sentences)

| Closeness | Sentences |
|---|---|
| 0.3571 | 1 |
| 0.3704 | 3 |
| 0.3846 | 2 |
| 0.4000 | 5 |
| 0.4167 | 14 |
| 0.4348 | 17 |
| 0.4545 | 25 |
| 0.4762 | 25 |
| 0.5000 | 33 |
| 0.5263 | 24 |
| 0.5556 | 11 |

Table 5: the different values of closeness centrality and the number of sentences which have the same closeness centrality value (20-word sentences)

| Closeness | Sentences | Closeness | Sentences |
|---|---|---|---|
| 0.2353 | 1 | 0.3226 | 2 |
| 0.2381 | 1 | 0.3333 | 2 |
| 0.2564 | 1 | 0.3390 | 3 |
| 0.2632 | 1 | 0.3448 | 2 |
| 0.2703 | 3 | 0.3509 | 2 |
| 0.2740 | 1 | 0.3571 | 5 |
| 0.2857 | 1 | 0.3636 | 1 |
| 0.2941 | 4 | 0.3704 | 6 |
| 0.2985 | 2 | 0.3774 | 4 |
| 0.3030 | 4 | 0.3846 | 7 |
| 0.3077 | 1 | 0.3922 | 5 |
| 0.3125 | 2 | 0.4000 | 2 |
| 0.3175 | 6 | 0.4082 | 2 |

As is the case in degree centralities, sentences in different genres show different distributions of closeness centralities. As for the closeness centralities of 10-word sentences, Fiction has 33 10-word sentences with the closeness centrality 0.384 out of 160 (about 20%), while Journal has 2 10-word sentences with the same closeness centrality out of 28 (about 10%). These results suggest that sentences in Journal tend to be more embedded than those in Fiction. As for the closeness centralities of 20-word sentences, there is not the same value of closeness centrality which is found in Fiction and Journal.

### 3.3 Discussion

The distributions of degree centralities and that of closeness centralities suggest that the degree centrality and the closeness centrality of a sentence are dependent on the word count of the sentence; the increase in word counts results in smaller degree centralities and closeness centralities, and it also results in more diverse values of closeness centralities. However, if we control the word count of the sentences taken from different genres, we can make explicit that the difference in genre is reflected on the number of sentences of the same degree centrality and of the same closeness centrality. In order to have a broader understanding of their distributions, we need to explore the difference of them across different word counts and different genres, which will be the research topic in future.

## 4 Conclusion

This study demonstrated that different distributions of degree and closeness centralities are found among different genres in MASC500k, yet it is suggested that this result can be due to the fact that texts with shorter sentences tend to have larger degree centralities and closeness centralities. It is also suggested that degree and closeness centrality should be treated with precaution. The distributions of degree centralities and closeness centralities of the sentences with the same word count taken from difference genres of corpus seemed to reflect the difference of these genres. Further examinations of the degree centralities and closeness centralities of the sentences in a variety of genres will enable us to have a broader understanding of their distributions, which will be addressed in future research.

**References**

Argamon, S., M. Koppel, and G. Avneri. (1998). Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems.*

Beauchamp, M.A. (1965). An improved index of centrality. *Behavioral Science. 10*, 161-163.

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge, UK: Cambridge University Press.

De Marneffe, M.C., B. MacCartney and C. D. Manning. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC* 2006.

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 June 27, 2009, from http://nlp.stanford.edu/pubs/dependencies-coling08.pdf.

De Marneffe, M.C. and C. D. Manning. (2012). *Stanford Typed Dependency Manual revised for Stanford Parser v.2.0.4*. Retrieved Novermber 13, 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. (1978). Centrality in social networks. *Social Networks* vol.1, 215-239.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1-76.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In A.Marantz, Y. Miyashita, & W. O'Neil (Eds.). *Image, Language, Brain*. 95-126. Cambridge, MA: MIT Press.

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

Ide, N., and K. Suderman. 2004. The American National Corpus First Release. *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, 1681-84.

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. 2011. Syntactic Dependency Distance as Sentence Complexity Measure. *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics,* 313-316.

Sabidussi, G. (1966). Then centrality index of a graph. *Psychometrika.* 31, 581-603.

Santini, M. (2004, January). A shallow approach to syntactic feature extraction for genre classification. In *proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (pp. 6-7). Birmingham, UK.

Sleator, D. and Temperley, D. (1991). Parsing English with a Link Grammar. *Technical Report CMU-CS-91-196*, CMU.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition* 105, 300-333.

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

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