# Syntactic Dependency Distance as Sentence Complexity Measure

Masanori Oya

Mejiro University

masanorioya2008@gmail.jp

## Abstract

This study introduces the possibility of using the average dependency distances (ADDs) of a sentence as one of the unique measures to indicate its complexity. The ADD of a sentence is automatically acquired from the parsing output of three different sentence groups. The differences in the results are discussed.

## Keywords

Dependency syntax, dependency distance, sentence complexity, Dependency Locality Theory

## Introduction

The graph-theory based approach to calculate sentence complexity proposed by Oya (2010) was intended as an alternative of T-Unit analysis, but did not take into consideration the distance of dependency, which can also indicate the complexity of a sentence. Dependency Locality Theory (DLT) (Gibson 1998, 2000) proposes that longer dependencies require more efforts to process the sentences. Based on this insight on dependency length, Temperley (2006) conducts a corpus study on written English, and shows that different syntactic contexts shows different dependency-distance preferences.

This study implements the insights of their study into the issue of calculating the sentence complexity, as part of the effort to construct an automatic evaluation of the essays written by Japanese learners of English. This study focuses on the average dependency distance (henceforth ADD) of each sentence taken from three different sentence sets (a high school textbook used in Japan, essays written by Japanese learners of English, and sentences chosen randomly from a newspaper for linguistic research) and shows the differences and similarities in the ADDs among these sentence sets. It will be shown that the word count of a sentence and dependency length are weakly correlated with each other; that is, sentences with more words tend to be more complex in terms of ADD, but not necessarily, and sentences with the same word count can have different sentence complexity. It will also be shown that the differences among these sentence sets in terms of the ADDs of the sentences shorter than 10 words are not statistically significant, and the differences in the ADDs of the sentences with 20 words and over, and less than 30 are not as statistically significant as those with 10 words and over, and less than 20 words.

## 1 Previous study

### 1.1 Graph-centrality based approach

Oya (2010) proposed that the dependency relations among words in sentences can be represented as directed acyclic graphs (DAGs), and their structural properties such as flatness (degree centrality, or the degree of how many words depend on one word) and embeddedness (closeness centrality, or the degree of how many words there are between the main verb and a given word) can be calculated automatically in order to use them as complexity measures of these sentences.

Oya (2010) argues that centrality measures acquired from the DAG representation of a sentence is better than Minimal Terminable Units (T-Units; originally proposed in Hunt (1965), with many other definitions so far) and D-Level Scale (Rosenberg & Abbeduto (1987), Covington et al. (2006)), in that graph-centrality measures take into consideration the width (how many words depend on one word) and depth (how many words between the main verb and a given word) of dependency among words, which T-Unit approaches do not.

Another advantage of using graph centralities as complexity measures of sentences is that they are well-defined, and often used in the field of network analysis, and it is easy to acquire them automatically, provided that we have well-formatted data.

### 1.2 Dependency Distance

The drawback of these centrality measures is that they abstract away the linear order of words in a sentence, hence they do not show the dependency distance between a head and its dependent. For

example, the DAG representation for the sentence "Sarah read the book quickly and understood it correctly" is as follows:

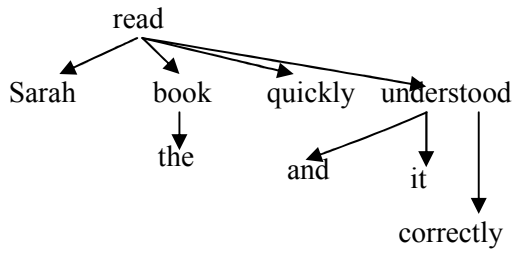


*Figure 1: The DAG representation for "Sarah read the book quickly, and understood it correctly."*

The DAG representation in Figure 4 does not preserve the surface order of the words in a sentence. This representation does not pose any problem in calculating degree centrality and closeness centrality; however, it is natural to consider that the word order of a sentence is the essential part of its syntactic property, hence it can be as relevant to syntactic complexity of sentences as centrality measures are.

The surface word order of a sentence highlights the dependency distance between a head and its dependent. Consider Figure 2 below:

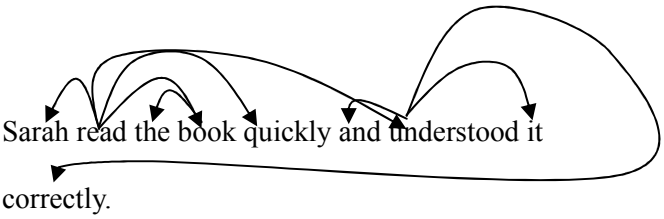


*Figure 2: The DAG representation for a sentence "Sarah read the book quickly and understood correctly.", preserving the word order*

The dependency relationships between the heads and tails are the same in Figure 1 and 2, yet the dependency distances between them are preserved in Figure 2; for example, this representation preserves that information that the head "read" is five words away from one of its dependent "understood".

### 1.3 Dependency Locality Theory (DLT)

Dependency Locality Theory (DLT) (Gibson 1998, 2000) proposes that the syntactic complexity of sentences increases in proportion to the length of syntactic dependency, and that the syntactic complexity of a sentence can be predicted by two factors: ''storage cost'' and "integration cost". Storage cost is that of keeping the previous words in memory. Integration cost is that of connecting the words in memory. Longer dependency lengths require more storage cost, increasing the difficulty of processing the dependency.

Temperley (2006) points out that the insight of DLT can be applied to the production of sentences; that is, he proposed that there are differences in preference for longer or shorter dependency lengths with respect to different syntactic environments.

Among other syntactic environments, he shows that subject NPs in S-V order quotations tend to be shorter than subject NPs in V-S order quotations (Temperley 2006: 307). Consider the examples below:

(1)
a. "I've read this book", Sarah said.
b. "I've read this book", said Sarah.
c. "I've read this book", my supervisor said.
d. "I've read this book", said my supervisor.

The subject NP in an S-V order quotation in (1a) is shorter than that in a V-S order quotation in (1d). The dependency length of the main verb "said" and its indirect-speech complement is shorter in (1d) than that in (1c), and (1d) is preferred to (1c).

Temperley's (2006) observation on the difference of dependency-length preference according to syntactic environments can be applied to the objective of calculating sentence complexity. For example, if Japanese learners of English prefer less complex sentences to more complex ones when they produce English sentences, the ADD of their written productions will be shorter than that of native speakers'. Thus, the ADD of a given text reflects the writers' preference on sentence complexity.

## 2 Analyses

In order to verify the insight proposed in the last section, corpus-based analyses are conducted.

### 2.1 Procedure

The ADD of the sentences in a text is calculated in the procedure summarized as follows:

Step 1: Parse the text by Stanford Parser
Step 2: Calculate the ADD from the parser output
Step 3: Analyze the data statistically

#### 2.1.1 Parsing text by Stanford Parser

I use in this study Stanford Parser (De Marneffe & Manning (2010)), which is a state-of-the-art dependency parser. It outputs the typed-dependency relations among the words in an input sentence in the following format:

(2)
nsubj(read-2, Sarah-1)
det(book-4, the-3)
dobj(read-2, book-4)

advmod(read-2, quickly-5)
cc(read-2, and-6)
conj(read-2, understood-7)
dobj(understood-7, it-8)
advmod(understood-7, correctly-9)

The triples in (2) are the output for the sentence "Sarah read the book quickly and understood it correctly." The first line "nsubj(read-2, Sarah-1)" indicates that the second word in the sentence has a dependent word "Sarah", which is the first word of the sentence, and the dependency type of this dependency is nsubj, or nominal subject.

#### 2.1.2 Calculating the ADD

The format of the output of Stanford Parser enables us to calculate the ADD easily. The triple "nsubj(read-2, Sarah-1), for example, shows that the distance of the dependency between these words is 2 – 1 = 1.

The ADD of a sentence is the sum of the distance of all the dependencies in the sentence divided by the number of the dependencies of the sentence. For example, the ADD of the sentence "Sarah read the book quickly and understood it correctly" is 19 / 8 = 2.735. The ADD of a text is the sum of the ADD of the sentences in the text divided by the number of sentences in the text. Along with the sentence-level calculation, we can acquire from the output of Stanford Parser the ADD of each of the dependency types.

#### 2.1.3 Statistical analyses

The ADDs of the sentences acquired from different text can be statistically analyzed. If we assume that the sentences written by native speakers of English are more complex than those written by Japanese learners of English, it can be expected that the ADD of the sentences written by native speakers of English is longer than those written by Japanese learners of English. Analysis of Variance (ANOVA) enables us to verify these expectations.

It can be expected that the number of words in a sentence has some effect on the ADD; the ADD of a sentence can increase in proportion to the word count. In order to take this effect into consideration, sentences of a certain range of word count are chosen from each of the text, and compare the ADDs.

### 2.2 Description of the text data

This study uses the following three different sets of sentence data:

1) Sentences taken from an English textbook used in Japanese high schools; 396 sentences in total (henceforth *Textbook*)
2) Sentences taken from the essays in English on the same topics ("self-introduction" and "happiness factors") written by Japanese learners of English (data used in Yoshida et al. (2009) and Oya (2010); 342 sentences in total (henceforth *Japanese*)
3) Sentences in the Parc 700 Dependency Bank, which are randomly extracted from section 23 of the UPenn Wall Street Journal Treebank; 676 sentences in total (henceforth *Parc*)

## 3 Results

### 3.1 Sentence-level comparison

Table 1 shows the descriptive statistics of the ADD of sentences in these three texts:

Table 1: The descriptive statistics of the ADDs

| | **Average** | **SD** | **Var** |
|---|---|---|---|
| *Textbook* | 2.075 | 0.599 | 0.358 |
| *Japanese* | 2.437 | 0.728 | 0.531 |
| *Parc* | 2.519 | 0.668 | 0.445 |

Table 2 shows the correlation coefficients of the ADD of sentences to their word count:

Table 2: The correlation coefficients of the ADDs to word counts

| | ***r*** |
|---|---|
| *Textbook* | .693 |
| *Japanese* | .670 |
| *Parc* | .528 |

Table 3 shows the result of ANOVA on the ADDs in all the sentences in these data; the difference in the ADD of all the sentences in these three texts is statistically significant.

Table 3: Analysis of Variance among Textbook, Japanese and Parc; **p<0.01

| **S.V** | **SS** | **DF** | **MS** | **F** | |
|---|---|---|---|---|---|
| **A** | 51.022 | 2 | 25.511 | 57.74 | ** |
| **Sub** | 623.884 | 1412 | 0.442 | | |
| **Total** | 674.906 | 1414 | | | |

Table 4 shows the result of ANOVA on the same data, but only the ADDs of the sentences with less than 10 words are analyzed.

Table 4: Analysis of Variance among Textbook, Japanese and Parc; word count <10

| **S.V** | **SS** | **DF** | **MS** | **F** | |
|---|---|---|---|---|---|
| **A** | 1.008 | 2 | 0.504 | 2.32 | n.s. |
| **Sub** | 50.99 | 235 | 0.217 | | |
| **Total** | 51.998 | 237 | | | |

Table 5 shows the result of ANOVA of the ADDs of the sentences with 10 words and over, and less than 20 words:

Table 5: Analysis of Variance among Textbook, Japanese and Parc; 10<= word count <20; **p<0.01

| S.V | SS | DF | MS | F | |
|---|---|---|---|---|---|
| **A** | 3.3381 | 2 | 1.6691 | 6.33 | ** |
| **Sub** | 151.3339 | 574 | 0.2636 | | |
| **Total** | 154.6720 | 576 | | | |

Table 6 shows the result of ANOVA of the ADDs of the sentences with 20 words and over, and less than 30 words:

Table 6: Analysis of Variance among Textbook, Japanese and Parc; 20<= word count <30; +p<.10

| S.V | SS | DF | MS | F | |
|---|---|---|---|---|---|
| **A** | 1.7842 | 2 | 0.8921 | 2.85 | + |
| **Sub** | 120.6673 | 386 | 0.3126 | | |
| **Total** | 122.4515 | 388 | | | |

## 4 Discussion

As is expected, the ADD of the sentences in the text Parc, which were written by native speakers of English, is the longest among these sentences.

The lower correlation between the word count and the ADD in the text Parc suggests that longer sentences do not necessarily have longer dependency distances.

Analyses of Variance on the ADDs of sentences with different word counts show that the ADDs of sentences with 10 words and over and less than 20 words are statistically different. This suggests that ADDs of short sentences (with word counts of less than 10 words), and long sentences (with word count of 20 words and over) do not serve as measures of sentence complexity as distinctively as those of sentences with word counts of 10 and over, and less than 20.

## 5 Conclusion

This study introduced the possibility of using the ADDs of a sentence as one of the unique measures to indicate its complexity. The ADD of a sentence is automatically acquired from the parsing output of three different sentence groups. The differences in the results are discussed.

## References

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). How complex is that Sentence? a proposed revision of the Rosenberg and Abbeduto D-Level Scale. CASPR Research Report 2006-01.

De Marneffe, M.C. & Manning, C. (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. & Manning, C. (2010). Stanford Typed Dependency Manual. Retrieved Retrieved July 3, 2010, from http://nlp.stanford.edu/software/dependencies_manual.pdf

Freeman, L. (1979). 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* (pp. 95–126). Cambridge, MA: MIT Press.

Hunt, K.W. (1965). *Grammatical Structures Written at Three Grade Levels*. Champaign, IL: National Council of Teachers of English.

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

Oya, M. (2010). Directed Acyclic Graph Representation of Grammatical Knowledge and its Application for Calculating Sentence Complexity. *Proceedings of the 15th Conference of Pan-Pacific Association of Applied Linguistics*.

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8, 19-32.

Scott, J.(ed.). (2002). Social Networks vol.1. London: Routridge.

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

Wasserman, S. & Faust, K. (1994). *Social Network Analysis*. Cambridge: Cambridge University Press.