# Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data

**Saeko Komori**
Chubu University / JAPAN
komori@isc.chubu.ac.jp

**Masatoshi Sugiura**
Nagoya University / JAPAN
sugiura@nagoya-u.jp

**Wenping Li**
Dalian Maritime University / CHINA
lwplovely1023@gmail.com

## Abstract

The purpose of this study is to examine methods of measuring syntactic complexity by analyzing an original corpus of written Japanese data from native speakers and learners of Japanese. We compared two measures, mean dependency distance (MDD) and mean hierarchical distance (MHD), which have been examined using in English in previous studies. Our research question is to compare the two methods and evaluate them in order to develop an index for measuring Japanese learner's syntactic complexity.

## 1 Introduction

Ortega (2015) overviewed recent SLA writing and syntactic complexity studies and discussed the reasons for inconclusive results among the studies. She observed that there are some factors that might affect differences in results across studies. One of them is a factor of measurements, and three measurements were discussed: 1) Subordination measures, 2) Length-based measures, and 3) Frequency-based measures. We believe that this factor needs to be studied further, and more precise indexes are necessary to measure syntactic complexity. This paper will examine mean dependency distance (MDD) and mean hierarchical distance (MHD) as good candidates for measuring L2 development of Japanese syntactic complexity.

## 2 Previous Studies on MDD and MHD

We will first review five studies using MDD and MHD as measures for syntactic complexity. Three of them were studies using native speaker (NS) data, and two used non-native speaker (NNS) data as summarized in Table 1.

| Study | MDD/MHD | Language | NS/NNS |
|---|---|---|---|
| Jing and Liu (2015) | MDD and MHD | English and Czech | NS |
| Jing and Liu (2016) | MHD and other measures | English | NS |
| Liu et al. (2017) | MDD | 20 natural languages | NS |
| Ouyang and Jiang (2017) | MDD | English | NNS |
| Komori et al. (2018, 2019) | MDD and MHD | Japanese | NNS |

Table 1: Summary of previous studies of the MDD and MHD

First, Jing and Liu (2015) studied both MDD and MHD using English and Czech as the first language. In order to examine the structural complexity of language, they compared two SVO languages: English with rigid word order and Czech with relatively free word order. They reported significant positive correlations between sentence lengths (SL), MDD, and MHD. They also discovered that "for longer sentences, English prefers to increase the MDD, while Czech tends to enhance the MHD" (Jing and Liu 2015, 161).

Second, the purpose of Jing and Liu (2016) was to analyze the hierarchical structure of English sentences, and they examined several different measures, including the MHD using a large English dependency treebank. As a result, they found significant positive correlations between the Vertices number (VN), the Hierarchical number (HN) and the MHD.

Third, Liu et al. (2017) was a cross-language examination of the MDD using 20 natural languages. They posited that dependency distance minimization is probably a universal regularity in human languages (Liu et al. 2017, 176).

Fourth, Ouyang and Jiang (2017) adopted the same calculation method as Liu et al. (2017) in order to examine if the MDD works as a measure of the language proficiency of second language learners. They conducted a study using Chinese EFL learners' compositions in eight grades from the first year of junior high school to the second year of university and reported the MDD increase from 1.845 in the first year of junior high school to 2.466 in the second year of university (Jiang and Ouyang 2017, 210). This results showed that the MDD could indicate the syntactic complexity of the learners' English. Jiang and Ouyang (2017) reported that the MDD measured sentence difficulty and how the MDD changed with the increase of learners' language proficiency across their learning levels.

Lastly, Komori et al. (2018 and 2019) examined the MDD and MHD with Chinese L1 learners of Japanese using Yokohama National University corpus (YNU, Kanazawa, ed., 2014). The learners in the YNU were all advanced learners, and were further divided into three levels: high (H), mid (M), low (L). As a result, there was not a significant difference in the MDD among the three levels of advanced learners. A gradual increase from L to H in the MHD, on the other hand, was found as their levels progressed as shown in Table 2.

| Group | MDD | MHD | Words | Number of Sentences |
|---|---|---|---|---|
| L | 2.16 | 1.75 | 8,806 | 1,316 |
| M | 2.08 | 1.84 | 10,525 | 1,523 |
| H | 2.16 | 1.98 | 10,810 | 1,391 |
| NS | 2.07 | 1.97 | 9,022 | 1,209 |

Table 2: MDD and MHD scores of YNU data

Komori et al. (2018 and 2019) examined advanced learners' syntactic complexity using the MDD and the MHD, but they examined only advanced learners. It is still unclear if the MDD and the MHD can measure language proficiency or language development. Therefore, in this study, we will examine if we can use the MDD and the MHD in order to measure Japanese learners' syntactic complexity using intermediate learners' corpus data. We also see if there are any differences between the two measures of the MDD and the MHD with intermediate learners' data to figure out what kind of differences the MDD and the MHD are measuring.

## 3 The Current Study

In order to examine the MDD and the MHD as syntactic complexity measures with Japanese learners, we collected our original written data from both learners and native speakers of Japanese. The following will describe the methods and materials of this study.

### 3.1 Participants

We started the data collection in 2018 with the aim to analyze learners' syntactic development. We collected written data and observed their development over time as their learning progressed. We asked each participant to write an argumentative essay on a manuscript paper of more than 600 characters without referring to any dictionaries. For native speakers, there was a time limit of 30 minutes, but the learners had 50 minutes to write an essay. The university students who participated in this project were the second (C2) and third-year (C3) university students. They were all Chinese native speakers majoring in Japanese in China. We analyzed the data from the intermediate level learners as well as Japanese native speakers (JP). For this particular study, there are 38 C2, 33 C3, and 35 JP compositions for comparison.

### 3.2 Corpus Data

We manually input each hand-written composition into the computer to compile corpus data. Table 3 shows the outline of the current corpus data. The topic of the composition used for the current study is "Will you decide your plans for life after graduation by yourself or will you consult other people?" which was in Japanese.

| Group | Participants | Sentences | Type | Token |
|---|---|---|---|---|
| C2 (second year university learners) | 38 | 721 | 1,269 | 10,296 |
| C3 (third year university learners) | 33 | 605 | 1,519 | 11,786 |
| JP (Japanese university students) | 35 | 463 | 1,462 | 12,495 |

Table 3: Outline of the current corpus data

After the data collection, we excluded outlier sentences with less than 4 words and also more than the number of the upper limit, which is upper quartile plus 1.5 interquartile range of the data in each group. As a result, we eliminated 129 (18%), 56(9%), and 34 (7%) of C2, C3, and JP outliers from the data, respectively.

### 3.3 Analysis

To parse the data, we formatted each composition to one sentence per line. Then, each sentence was parsed syntactically with Cabocha, a Japanese dependency structure analyzer (Kudo and Matsumoto, 2002) and IPADic, and the data was edited by retrieving dependent ID, governor ID and the original word as illustrated in Table 4. After editing, we used the dependent ID and governor ID to calculate the dependency distance (DD), the difference between governor ID and dependent ID. Then, we used the following two formulas (1) and (2) to calculate the MDD of a sentence or text, according to Liu et al. (2017). Finally, we used the dependent ID and governor ID to construct dependency trees and calculated the MHD for each sentence with Python scripts, as shown in Figure 1.

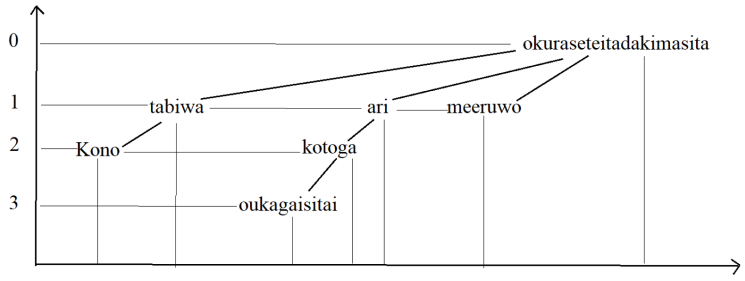
| Dependent | | Governor | Dependent ID | Governor ID | DD | HD |
|---|---|---|---|---|---|---|
| Kono | => | tabiwa | 0 | 1 | 1 | 2 |
| tabiwa | => | okuraseteitadakimasita | 1 | 6 | 5 | 1 |
| oukagaisitai | => | kotoga | 2 | 3 | 1 | 3 |
| kotoga | => | ari | 3 | 4 | 1 | 2 |
| ari | => | okuraseteitadakimasita | 4 | 6 | 2 | 1 |
| meeruwo | => | okuraseteitadakimasita | 5 | 6 | 1 | 1 |

Table 4: Method of calculating DD and HD

$$MDD(\text{the sentence}) = \frac{1}{n-1}\sum_{i=1}^{n}|DD_i| \quad (1)$$

$$MDD(\text{the text}) = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \quad (2)$$

In formula (1), $n$ is the number of words in the sentence, and $DD_i$ is the DD of the $i$-th syntactic link of the sentence. In formula (2), $n$ is the total number of words in the text, $s$ is the total number of sentences in the text.




HD = 2 + 1 + 3 + 2 + 1 + 1
MHD = HD / (V - 1)
= 10 / 6
= 1.67

Figure 1: MHD calculation

### 3.4 Results

Our analysis shows that both the MDD and the MHD increased from C2 to C3 as is shown in Table 5. This means that the increase may reflect their syntactic complexity development as their Japanese learning progressed.

| | | | Median | |
|---|---|---|---|---|
| Group | Number of Sentences | SL (Min, Max) | MDD (Min, Max) | MHD (Min, Max) |
| C2 | 592 | 6 | 1.91 | 1.67 |
| | | (4, 4) | (1.00, 4.00) | (1.00, 4.00) |
| C3 | 547 | 8 | 2.00 | 2.00 |
| | | (4, 18) | (1.00, 4.21) | (1.00, 4.64) |
| JP | 429 | 10 | 2.00 | 2.50 |
| | | (4, 24) | (1.00, 3.96) | (1.00, 8.17) |

Table 5: SL, MDD and MHD comparison of C2, C3 and JP

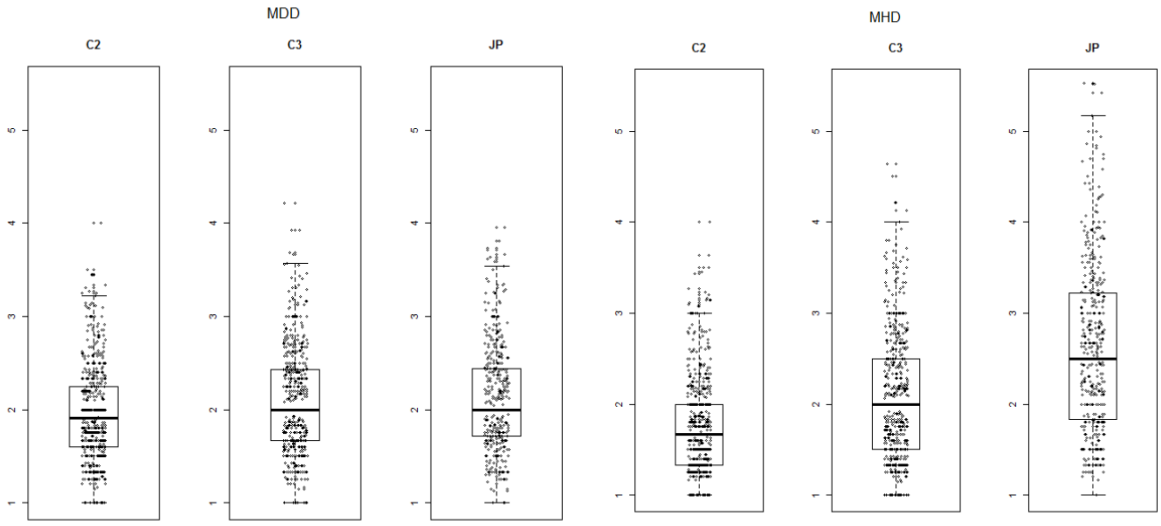



Figure 2: Boxplots with jitter of the MDD and the MHD for C2, C3 and JP

Figure 2 shows the boxplots of the MDD (on the left) and MHD (on the right). It is easy to see a gradual increase of score from C2 to C3 to JP for the MHD. Non-parametric statistical analyses of multiple comparisons were conducted. Table 6 shows Brunner-Munzel (BM) Test results as well as effect sizes (Cliff's delta).

| | MDD | | | MHD | | |
|---|---|---|---|---|---|---|
| | BM | *p* | Cliff's delta | BM | *p* | Cliff's delta |
| C2 v C3 | 3.88 | .0001 | .13 (negligible) | 7.73 | <.0001 | .25 (small) |
| C3 v JP | 1.04 | .2988 | .04 (negligible) | 10.26 | <.0001 | .35 (medium) |
| C2 v JP | 4.86 | <.0001 | .17 (small) | 19.22 | <.0001 | .56 (large) |

Table 6: Brunner-Munzel Test and Cliff's delta of the MDD and MHD

The results of the analyses along with the interpretation of effect sizes indicated that the MHD scores demonstrated significant group differences but the MDD scores did not. There was only a small difference between C2 and JP, but no other significant group differences were observed in the MDD scores. As for the MHD, on the other hand, significant increases can be observed. From our current data, we may conclude that intermediate Japanese learners' syntactic complexity increased in terms of the MHD, but it is difficult to conclude that the MDD showed any increase.

Figure 3 shows correlations between sentence length (SL), MDD, and MHD. They are all significantly correlated ($p < 0.01$). The correlation coefficients between SL and MHD in JP are highest (0.72), and those in C3 and C2 are also moderate (0.67 and 0.62). Correlations between MDD and MHD are not observed in any of the three groups. It can be interpreted that both MDD and MHD are measuring syntactic complexity, but they do not measure the same complexity. Further study is necessary to uncover what the differences are between the syntactic complexities measured by the MDD and the MHD.

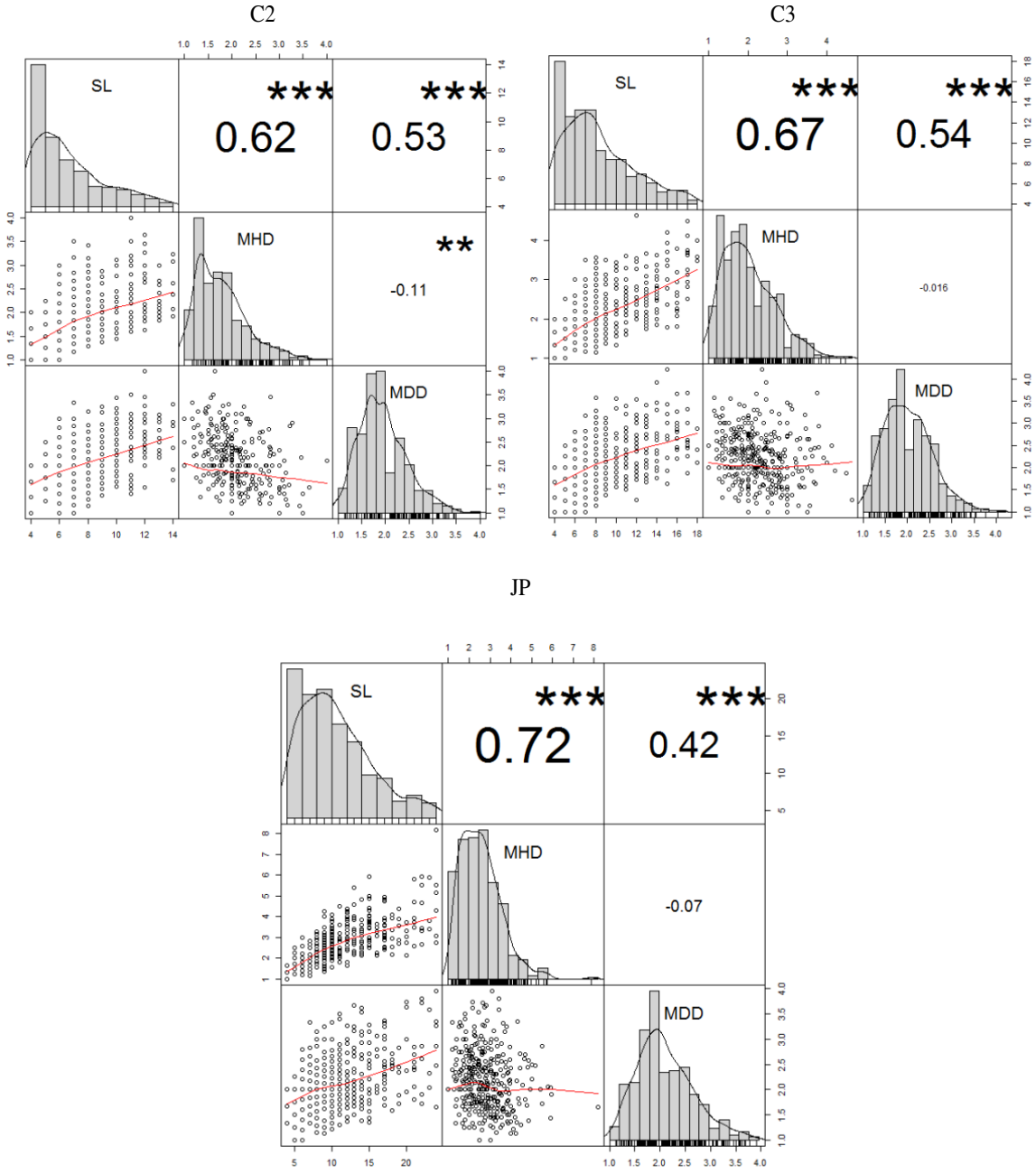



Figure 3: Correlations between SL, MDD and MHD of C2, C3 and JP

## 4 Discussion

From our data analyses of the intermediate learners and native speakers of Japanese, we showed that Japanese learners' syntactic complexity can be measured with the MHD, but it is not as clear with the MDD. As for the learners' proficiency levels, learners in C2 and C3 of the current study were intermediate learners who studied Japanese for about 13 months

(C2) and 24 months (C3) in China, whereas participants in YNU data in Komori et al. (2018, 2019) were all living in Japan and had studied 20 months to16 years. The MDD from the YNU learners did not show any increase, which may indicate that they might have reached a plateau period. The MDD scores of the intermediate learners in this current study show some increase between groups (C2 and C3), but it is not statistically significant and its effect size is negligible, thus MDD may not denote learners' syntactic development. As for the MHD, the previous study also showed an increase even among advanced learners. In this respect, the MHD might be a better measure to show Japanese learners' syntactic development for both intermediate and advanced learners (Komori, et al., 2019). There may be some linguistic preferences between the MDD and the MHD in Japanese, as is discussed in Jing and Liu (2015) with English and Czech for longer sentences. It may also be argued that some of the characteristics of Japanese syntactic complexity appeared with MHD rather than MDD. As for the composition in terms of genre, the current study used argumentative essays which may contain relatively longer sentences, while the data in YNU consist of 12 different topics and they include short email messages as well (Kanazawa ed. 2014). These two factors (level of learners and genre) may have influenced the results, which we need to control in future studies.

As we have seen above, the MHD may be used to measure learners' syntactic development, but we need to further scrutinize and define the MDD and the MHD as syntactic complexity measures. There are also some problems to be solved in future studies. First of all, the learners' compositions contain errors, and they may cause analytical errors of syntactic complexity. There is also a matter of genre. We only analyzed one topic of compositions in the current study. We are planning to collect compositions with several different topics. Finally, a longitudinal study is necessary to examine the learners' development over time.

## Acknowledgements

This study is supported by JSPS KAKENHI Grant Number JP19K00749 and JSPS KAKENHI Grant Number 16H03444.

## References

Haitao Liu, Chunshan Xu and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages, *Physics of Life Reviews*, 21, 171-193.

Jinghui Ouyang and Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguisitics,* October 2017, 1-20.

Jingyang Jiang and Jinghui Ouyang. 2017. Dependency distance: A new perspective on the syntactic development in second language acquisition Comment on "Dependency distance: A new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews* 21, 209-210.

Hiroyuki Kanazawa ed. 2014 *Nihongo kyoiku no tame no tasuku betsu kakikotoba kopasu* (Corpus of task-based writing for Japanese language education), Hitsuji, Tokyo.

Lourdes Ortega. 2015. Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing,* 29, 82–94.

Saeko Komori, Masatoshi Sugiura and Wenping Li. 2018. Examining the applicability of the mean dependency distance (MDD) for SLA:A case study of Chinese learners of Japanese as a second language. *Proceedings of the 4th Asia Pacific Corpus Linguistic Conference (APCLC 2018),* 237-239.

Saeko Komori, Masatoshi Sugiura and Wenping Li. 2019. Evaluating mean dependency distance (MDD) and mean Hierarchical distance (MHD) to measure development of Japanese syntactic complexity. *The 2019 conference of the American Association for Applied Linguistics (AAAL).*

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking*, CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002*, 63-69.

Yingqi Jing and Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. *Proceedings of the Third International Conference on Dependency Linguistics,* 161-170.

Yingqi Jing and Haitao Liu. 2016. A quantitative analysis of English hierarchical structure. *Journal of Foreign Languages,* 39, 2-11.