On Mean Dependency Distance as a Metric of Translation Quality Assessment

Chenliang Zhou [a, *]

[a] Department of Linguistics, Zhejiang University, Hangzhou, China.
*Corresponding author Email: zclsam@outlook.com

Abstract: This paper has adopted a quantitative approach to carry out a linguistic study, within the theoretical framework of dependency grammar. Translation is a process where source language and target language interact with each other. The present study aims at exploring the feasibility of mean dependency distance as a metric for automated translation quality assessment. The current research hypothesized that different levels of translation are significantly different in the aspect of mean dependency distance. Data of this study were based on the written translation in Parallel Corpus of Chinese EFL Learners which was composed of translations from Chinese EFL learners in various topic. The translations were human-scored to determine the levels of translation, according to which the translations were categorized. Our results indicated that: (1) senior students perform better in translation than junior students, and mean dependency distance of translations from senior group is significantly shorter than the junior; (2) high quality translations yield shorter mean dependency distance than the low quality translations; (3) mean dependency distance of translations is moderately correlated with the human score. The resultant implication suggests the potential for mean dependency distance in differentiating translations of different quality.

Keywords: Dependency Grammar, Mean Dependency Distance, Translation Quality Assessment, Quantitative Linguistics.

About the Author

Chenliang Zhou is a researcher in the School of International Studies, Zhejiang University. His research interests include quantitative linguistics and language evolution. He has given several reports on the topics of quantitative approaches to linguistic studies. He is currently interested in the internal and external factors to language evolution.

1. Introduction

Dependency distance, derived from dependency grammar, is the linear distance between two words which are in a dependency relation. Similar to Chomsky's transformational-generative grammar, dependency grammar also illustrates a method for sentence parsing. As previous studies (Fan & Jiang, 2019; Liang, Fang, Lv, & Liu, 2017; Wang & Liu, 2017) have demonstrated, mean dependency distance (hereinafter MDD), as one of the most prominent quantitative linguistic features in dependency grammar, is deemed as an insightful metric of syntactic complexity and has been widely investigated for its universal reliability in second language studies and language typology (Fan & Jiang, 2019). Syntactic complexity was conceptualized for measuring the comprehension difficulty of human language processing (Liu, 2008). However, it has been a challenge for a long time to find a suitable metric to quantify such complexity. Yngve (1960) proposed the depth of a parsed syntactic tree as the metric for language processing complexity. The underlying rationale is the max symbols stored in working memory during the construction of a

sentence. Previous studies about complexity metrics involve the investigation of dependency relations between linguistic units in the context. Thus, it is natural to map the complexity of dependency structure to the cognitive burden. Liu (2008) has proven that MDD can take the role of a reliable metric for comprehension difficulty based on a mass cross-language project. On the basis of twenty-language syntactic annotated treebanks, he demonstrated that MDD is universally reliable in evaluating syntactic complexity, while MDD of Chinese is found to be relatively longer (3.662) than that of English (2.543) (Liu, 2008).

Apart from proving MDD's universality in measuring comprehension complexity, Liu (2008) also implied that dependency distance minimization is general among various human languages. Dependency distance minimization could be probably caused by human cognitive constraints (Liu & Lin, 2018). Other studies, for instance, using MDD as an index in language typology (Liu, 2010) to probe human language production mechanism (Lu, Xu, & Liu, 2016), provide us a whole new perspective for linguistic research, which could lead the linguistics onto the road of scientification. Analogously, the quantitative approach is a must and also the potential trend for the scientification of translation quality assessment. As the translation market demands a more reliable quantitative evaluation method, machine evaluation, with its objective and robust result, has gradually become the cutting-edge in translation quality evaluation research. However, due to the lack of linguistic interpretability of evaluation results, the current automatic evaluation method is difficult to be further applied in the large-scale standardized examination. Therefore, the model of translation quality assessment based on quantitative linguistic characteristics has a broad prospect for practical use. Such a model takes efficiency and linguistic-interpretability into account, and could even benefit translation teaching, with both theoretical and practical value.

A reliable automated translation quality assessment system should be backed up with a suitable linguistic metric to quantify the translation quality. Given the fact that the MDD of Chinese texts is longer than that of English (Liu, 2008), English translations, as a "third code" (Fan & Jiang, 2019; Jiang, Ouyang, & Liu, 2019; Liang, Fang, Lv, & Liu, 2017; Liang, Lv, & Liu, 2019), would inevitably be influenced by its source language. Fan and Jiang (2019) proved that translated texts (English translated from Chinese) and target texts (English) are significantly different from each other in the aspect of dependency distance, and translated texts yield longer MDD than the target texts. It is probably influenced by the increased cognitive burden in dealing with Chinese (Fan & Jiang, 2019). Since the experienced translator would pursue "the authentic English", which means the translational language should be as close to the target language as possible.

Thus, we hypothesized that MDD of good translation should converge to that of the target language while diverge from the source language. We can reasonably expand the applicability of MDD as a metric for differentiating translation quality. Nonetheless, few prior research have investigated what constitutes a reliable criterion for translation quality evaluation. To fill the gap, this study aims to explore the potential of MDD as a metric of translation quality assessment to quantify and differentiate translation quality. Moreover, utilizing an index to represent translation quality is what "quantifying" means. Quantitative assessment is more stable and efficient, for "scoring the translation" is a concrete representation of translation quality. With increasing intercultural communications happen globally, the academia and the industry pay more and more attention to standardized language test, which requires the evaluation criterion that takes both efficiency and fairness into account. As the quantitative approach is much more objective and energy-saving than the traditional human scoring, a suitable metric for translation assessment is in broad demand. To fill the gap in both theory and application, the current paper has investigated the feasibility for MDD to measure translation quality.

2. Materials and Methods

2.1 Preliminaries: dependency grammar and MDD

Dependency grammar is best described as a particular perspective on the syntactic analysis (de Marneffe & Nivre, 2019). Dependency grammar and constituency grammar provide two dominating approaches to syntactic parsing (Fan & Jiang, 2019). The common core of all theoretical stances of dependency grammar is the agreement that syntactic structure consists of binary unequal syntactic relationships between words, called dependency relations. The theoretical core of dependency grammar is the dependency relation, which is defined by the syntactic structure of a sentence. It is assumed that words are attached to each other relying on this relationship, whereby

one word is the governor (head) and the other the dependent (Liu, 2008). In such an approach, the sentence could be analyzed into a tree diagram according to the dependency relations between each pair of words. (Hudson, 2010; Liu, 2009; Nivre, 2006; Tesnière, 2015) The dependency tree is constructed by nodes that represent words and labeled arcs that represent different types of dependency relations. Liu (2009) has concluded three core properties of a dependency relation: (1) it is a binary relation between two words; (2) it is almost asymmetrical, however, with one of the two units, called the dependent, complementing or modifying the other, called the governor, pointing from the governor to the dependent with an arrowed arc. (3) it is labeled and the type of a dependency relation is drawn in the half-plane above the arc linking the two units.

Based on these properties, a dependency tree of the sentence *little boys speak English fluently* is illustrated in figure 1. The dependency tree emphasizes the linguistic unit's functional role in a sentence. The noun **boys** fills the subject role of verb **speak**, with the noun **English** filling the object role. The adjective **little** fills the attributive modifier role of noun **boys**, and the adverb **fluently** takes the adverbial modifier of predicate **speak**. Types of dependencies is indicated by the labeled arcs. The number in the half-plane below the word sequence indicates the linear order or position of the word in the sentence, serving as the index of calculating MDD.
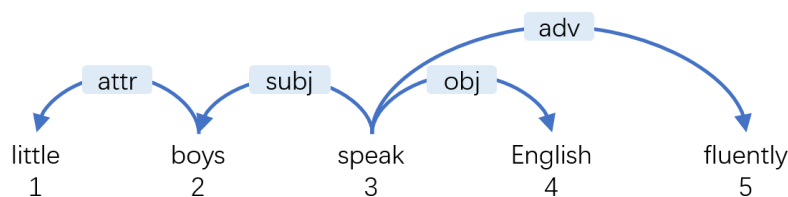

Figure 1. Dependency tree of the sample sentence "*little boys speak English fluently*".

The linear distance between the dependent and its governor is defined as dependency distance (Liu, 2009). In other words, dependency distance can be calculated by the difference between the linear order of the pair of words which are in a dependency relation. The concept was first used by Heringer, Strecker, and Wimmer (1980), and the term "dependency distance" was proposed by Hudson, who points out that DD is more compatible in dependency grammar than Phrase Structure Grammar (Fan & Jiang, 2019).

Liu, Hudson, and Feng (2009) have proposed a method for measuring the MDD of a sentence. The core of the method is the distance between the words pair in dependency relation. The numbers below the sentence, indicating the linear position of each word, are used for calculating MDD. Formally, let $W_1…W_i…W_n$ be a word string, with linear order as a subscript. For any dependency relation between two words, if W_a is a governor with W_b its dependent, then the dependency distance of this dependency relation can be defined as the difference a-b. By this measure, the DD of adjacent words is 1. When *a* is greater than *b*, the DD is a positive number, which indicates that the governor follows the dependent; when *a* is smaller than *b*, the DD is a negative number, which indicates that governor precedes the dependent. Howe ver, in measuring MDD, the current measure is the absolute value of dependency distance. The MDD of a sentence can be defined in formula (1):

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

In this formula, **n** is the sentence length and *DD_i* means the dependency distance of the **i-th** syntactic relation of the sentence. There is, in a sentence, generally one word, the root verb, without a governor. As a result, the DD of the root is defined as 0. Hence, in the sample sentence *little boys speak English fluently*, a sequence of |DDs| is 1 1 0 1 2, obtained by subtracting the order number of the word and that of its governor. Thus, we can tell that the MDD of the sample sentence is 5/4 = 1.25.

2.2 Materials

Parallel Corpus of Chinese EFL Learners (PACCEL) is employed by this study. PACCEL, with 2.1 million words in total, contains translations originally produced by junior and senior students of English major in eighteen Chinese

universities (Wen & Wang, 2008). Students were allowed to finish the translation in approximately 60 minutes without external help. Translations selected for the current study are extracted from PACCEL-W, a sub-corpus of PACCEL where 1.6-million-word written translations of texts on topics of literature, science, politics, economy, and social life. The source text for this study is an argumentative text of political topics with 320 Chinese characters. One thousand translations were randomly selected from the corpus, and translations were cleaned to delete the blank translations and convert Chinese punctuation marks into English ones for the convenience of computer processing.

2.3 Methods

To calculate the MDD of selected text more precisely, the current research utilized the automatic tool, leoDDcalculator, which is an R package calculating the values of MDD and NDD of texts (Lei & Jockers, 2020; Lei & Wen, 2019). The leoDDcalculator is based on UDPipe, an open-source tool for tokenization, tagging, and parsing of CoNLL-U files. The language model used in the present study is Universal Dependency 2.4, containing 90 models of 60 languages, and the precision for Part-of- Speech (PoS) annotation can reach 94%, which is efficient for the current study. Although leoDDcalculator can provide effective values of MDD, there still existed a few defects in the annotated data. Since the translations are the real-world performance of EFL learners, language mistakes in translations are also faithfully recorded, increasing the inaccuracy of the program (Ouyang & Jiang, 2018). As a result, preliminary parsing done by leoDDcalculator is manually checked and modified in this study.

2.4 Procedure

The study tries to explore the feasibility of MDD as a metric for quantitative translation quality assessment. This study was carried out in four stages (see figure 2.): translation selection; MDD calculation; human scoring and significant difference validation.
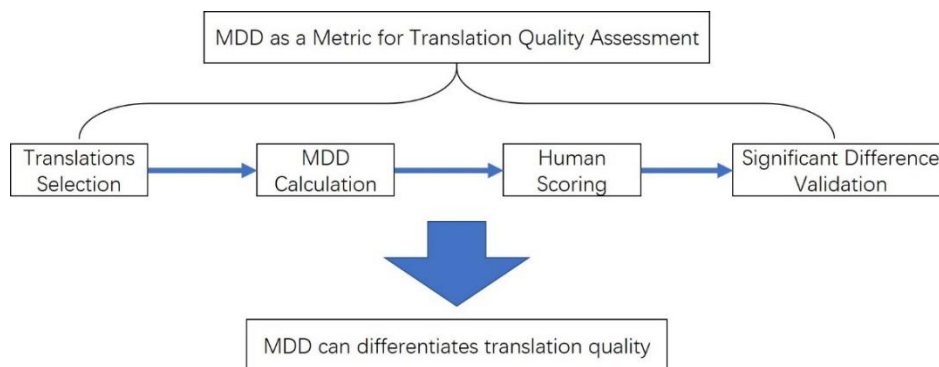

Figure 2. Research procedure for the current research

Selected translations are translated texts of a 320-Chinese-charterer argumentative text which was picked up to enhance the research performance, for the source text is relatively complex in the syntactic structure. Then, MDDs were automatically calculated by leoDDcalculator.

Human scoring was processed by three raters. They scored Chinese EFL learners' translations which we selected. All three raters are senior students who studied English literature and translation for 4 years. The raters all have attended translation courses and they have extensive experience in translation scoring. The average scores of all three raters were used to ensure interrater reliability.

Human scoring of this study followed the criterion for TEM-8 translation (Zou, 2013), concerning that TEM-8 is a required test for all who major in English and Translation in China. Three raters are familiar with this criterion. The scoring results of three raters endured statistical test to ensure the interrater reliability. The mean correlation coefficient between the scores assigned by three raters was 0.936, and Cronbach's alpha was 0.978, indicating a very high level of inter-rater reliability in this research.

The scale for TEM-8 translation scoring is 1–10, where score 1 is the worst translation with least faithfulness, and score 10 is the excellent translation with full faithfulness. However, human scoring for this research focused on the semantic and structural quality of the translation. As the aim of this study is to explore the feasibility of MDD on

discriminating translation quality, the human scoring was rated according to the translation's overall quality, i.e., the sentence structure. Hence, the minor mistakes in translations (e.g., misspell, inappropriate word choice) were not penalized, because the penalization process in translation quality assessment was out of the target of current research.

Finally, translated texts were grouped by their scores. The study performed *t*-test with the MDDs to examine the intergroup significant difference of different translation quality.

3. Results and Discussion

3.1 MDDs of translations in Chinese EFL learners

The fundamental purpose of this study is to validate the MDD's competence in differentiating translation quality. It is reasonable to carry out the study because the pre-mentioned perspective justifies the existence of translational language as a distinctive "third code" (Fan & Jiang, 2019). Following the previous studies (Fan & Jiang, 2019; Ouyang & Jiang, 2018), data of the present study reaches an agreement with those prior ones. Ouyang and Jiang (2018) confirmed that dependency distance distribution of EFL learners' written texts can measure their second language proficiency. In order to validate the hypothesis, we performed a *t*-test to examine whether the MDDs between different groups are statistically significant. The result indicated translations of junior students' group (JU, M = 2.256, SD = 0.234) yielded longer MDD than that of senior students' group (SE, M = 2.197, SD = 0.159), and there exists a significant difference in MDD between JU and SE (t = 2.879, *df* = 998, p < 0.05), see figure 3(a).

Moreover, the writer clustered translations into two groups according to human scores, one for a score above 6 (ABV6) another for the opposite (BLW6), for score 6 is in the middle of the scoring scale. In our corpus, 436 texts are categorized in group BLW6 and 564 texts in group ABV6, with BLW6 (M = 2.057, SD= 0.096) yielding longer MDD than ABV6 (M = 1.956, SD = 0.127). The result of a *t*-test with MDD demonstrated that ABV6 is significantly different from BLW6 in MDD (t = 3.015, *df* = 998, p < 0.05), see figure 3(b).
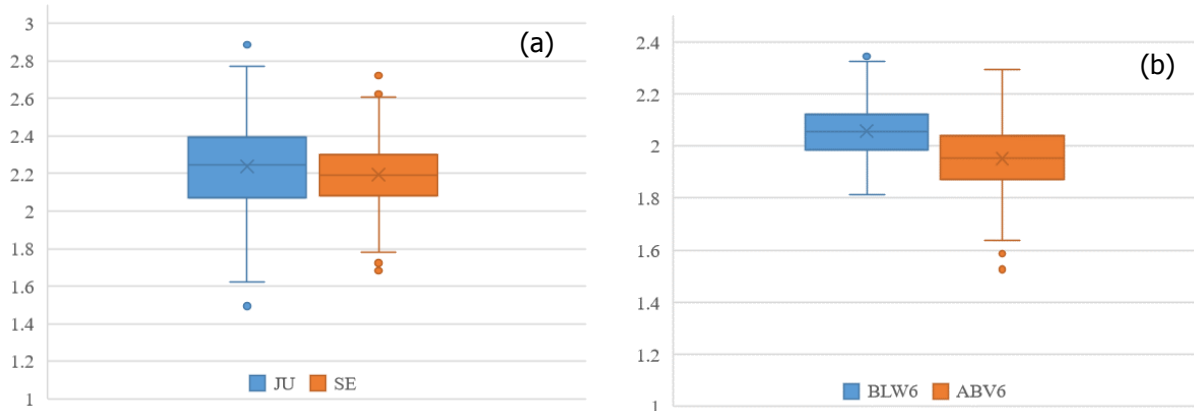

Figure 3. MDDs of group JU with SE (a), and group BLW6 with ABV6(b)

In order to further investigate the MDD's competence in classifying more fine-grained translation quality levels, the writer further performed a one-way ANOVA with MDD on each human score group (Group 1 for translations with human score 1, Group 2 for translations with human score 2, Group 3 for score 3, etc.) However, the result of a more fine-grained examination for dependency distance as metric of translation quality is not very satisfactory. The result indicated that the MDDs and the human scores are moderately negative correlated (Pearson's correlation coefficient = -0.378, p < 0.001). Moreover, the MDDs of each group did not show a significant difference among themselves, partially owing to the scarcity in human score data.

3.2 The Existence of Distinctive Translational Language

The result is legitimized by MDD of translational language (i.e., the English translated from Chinese, cf. Fan & Jiang 2019) turning out to be exactly between the source language (id. Mandarin Chinese) and target language

(id. English), and three types are significantly different from each other. Since previous studies have proven the MDD is an important measure for text's cognitive demands for processing the text, the fact that the MDD of translational language yields between the source language and target language implies the featured process of translation. There are two "opposing forces" confronting in the process of translating, as if source language and target language drag against each other, resulting in the distinctive translational language. Fan and Jiang (2019) demonstrated that a longer MDD of translational language represents a larger cognitive cost in processing than native languages, therefore, translational language is less authentic than native language and performing higher comprehension difficulty.

3.3 Natural Tendency of Dependency Distance Minimization

Other findings that could support the MDD's competence are corroborated by several corpus-based researches and psycholinguistic experiments, indicating that human language has a tendency toward dependency distance minimization (Ouyang & Jiang, 2018). This tendency suggests that human language share a universality of being "concise", although human languages differ in pronunciation, vocabulary, etc., at the syntax level there seems to have a universal constraint on cognitive mechanism. Our result indicated that the development of Chinese EFL learners' translation endured similar "minimization" process. Hence, the dependency distance minimization is considered as a result of human avoiding the use of long-distance dependency distances to reduce cognitive demands in language production and comprehension (Ouyang & Jiang, 2018). To conclude, lower MDD is somehow the nature of human language, and dependency distance minimization is driven by human cognitive constraint. In this sense, translations with lower MDD facilitate comprehension.

3.4 Principle of Least Effort and Translation

The principle of least effort, raised by Zipf (1949), stated that there is a pair of opposing forces (i.e., the force of unification and the force of diversification) that guides human beings' language production, and the present language results from the dynamic balance of two opposing forces (Liu, 2017). Since translation is a rather complex process of language transformation that contains both source language understanding and target language production, the "principle of least effort" should also take effect in translation.

The translator achieves the least effort by the simple equivalence of words between source language and target language, ignoring the higher hierarchies like syntax and function. As a result, such translation is mainly driven by the "force of unification" so that "word-by-word" translation is the extreme of this kind, i.e., translation is unified verbatim to the source language. However, such translation is hardly acceptable for the recipient, and recipients must make much effort to understand (decode) the translation. In contrast, the recipient achieves the least effort by requesting the translation equivalent on multiple hierarchies of language. As a result, such translation is mainly driven by the "force of diversification" so that native text with the exact equivalent meaning and form of source text is the extreme of this kind, which is a rather ideal "translation" type. To some extent, it is target language rather than translation. Under this circumstance, however, translators have to make much effort to produce this translation. The translation quality assessment should focus on whether recipients could easily understand the translation. In other words, it is how much effort recipients put in understanding that translation quality assessment model measures, with good translation requiring less effort in understanding. Therefore, good translations converge to the target language but diverge from the source language.

In the current research, both the translator and recipient want to achieve the least effort so that two opposing forces must ultimately reach a balance. Consequently, translators would seek a compromise between these two forces and create the translational language whose MDD fell between the source language and target language. Furthermore, the significant difference between different translation grade group reveals the development of Chinese EFL learners' translation proficiency.

4. Conclusion

The current study adopted multi-disciplinary approach to validate the feasibility of a novel metric for automated translation scoring system. This study has established the feasibility of MDD as a metric for assessing

translation quality. The present study assumes that the MDD of translation should converge to that of the target language and diverge from that of the source language. Furthermore, the research demonstrated that it is feasible to differentiate translation quality by MDD, for the translations with higher human scores basically yield shorter MDD and significant difference with MDD also exists between high human score group and low human score group. The present study also recognized a moderate correlation between MDDs and human scores.

The study could contribute to machine translation evaluation. It is more reliable for the assessment model to combine metrics with a linguistic basis than the pure computational method which is based on the statistical similarity between translation and reference text, for the latter approach was blamed for lacking linguistical interpretability. Moreover, the study may shed light on translation teaching, where MDD could be used as an indicator to trace the development of student's translation proficiency. Further research into automated translation quality assessment could adopt MDD as a metric and investigate to what extent can MDD reflects the translation quality. More systematic studies are required to reveal the macro level of translation quality, as statistical methods like multiple linear regression analysis could be used in further study to construct an automated translation quality assessment model.

References

de Marneffe, M. C., & Nivre, J. (2019). Dependency grammar, Annual Review of Linguistics, 5, 197-218. https://doi.org/10.1146/annurev-linguistics-011718-011842

Fan, L., & Jiang, Y. (2019). Can dependency distance and direction be used to differentiate translational language from native language?, Lingua, 224, 51–59. https://doi.org/10.1016/j.lingua.2019.03.004

Heringer, H. J., Strecker, B., & Wimmer, R. (1980). Syntax: Fragen, Lösungen, Alternativen, Fink.

Hudson, R. (2010). An introduction to word grammar, Cambridge University Press, Cambridge, United Kingdom. https://doi.org/10.1017/CBO9780511781964

Jiang, J., Ouyang, J., & Liu, H. (2019). Interlanguage: A perspective of quantitative linguistic typology, Language Sciences, 74, 85-97. https://doi.org/10.1016/j.langsci.2019.04.004

Lei, L., & Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure, Journal of Quantitative Linguistics, 27(1), 62-79. https://doi.org/10.1080/09296174.2018.1504615

Lei, L., & Wen, J. (2020). Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses, Lingua, 239, 102762. https://doi.org/10.1016/j.lingua.2019.102762

Liang, J., Fang, Y., Lv, Q., & Liu, H. (2017). Dependency distance differences across interpreting types: Implications for cognitive demand, Frontiers in Psychology, 8, 2132. https://doi.org/10.3389/fpsyg.2017.02132

Liang, J., Lv, Q., & Liu, Y. (2019). Quantifying interpreting types: Language sequence mirrors cognitive load minimization in interpreting tasks, Frontiers in psychology, 10, 285. https://doi.org/10.3389/fpsyg.2019.00285

Liu, H. T. (2008). Dependency distance as a metric of language comprehension difficulty, Journal of Cognitive Science, 9(2), 159-191. https://doi.org/10.17791/jcs.2008.9.2.159

Liu, H. T. (2009). Dependency Grammar: From Theory to Practice, Science Press, Beijing, China.

Liu, H. T. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks, Lingua, 120(6), 1567-1578. https://doi.org/10.1016/j.lingua.2009.10.001

Liu, H. T. (2017). An Introduction to Quantitative Linguistics, The Commercial Press, Beijing, China.

Liu, H. T., & Lin, Y. N. (2018). Methodology and trends of linguistic research in the era of big data, Journal of Xinjiang Normal University (Philosophy and Social Sciences), 39(1), 72-83.

Liu, H. T., Hudson, R. & Feng, Z. (2009). Using a Chinese treebank to measure dependency distance, Corpus Linguistics and Linguistic Theory, 5(2), 161-174. https://doi.org/10.1515/CLLT.2009.007

Lu, Q., Xu, C., & Liu, H. (2016). Can chunking reduce syntactic complexity of natural languages?, Complexity, 21(S2), 33-41. https://doi.org/10.1002/cplx.21779

Nivre, J. (2006). Inductive dependency parsing, Springer Netherlands. https://doi.org/10.1007/1-4020-4889-0_4

Ouyang, J., & Jiang, J. (2017). Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners?, Journal of Quantitative Linguistics, 1−19. https://doi.org/10.1080/09296174.2017.1373991

Tesnière, L. (2015). Elements of structural syntax, John Benjamins Publishing Company. https://doi.org/10.1075/z.185

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction, Language Sciences, *59*, 135-147. https://doi.org/10.1016/j.langsci.2016.09.006

Wen, Q. F. & Wang, J. Q. (2008). Parallel Corpus of Chinese EFL Learners, Foreign Language Teaching and Research Press, Beijing, China.

Yngve, V. H. (1960). A model and an hypothesis for language structure, Proceedings of the American philosophical society, 104(5), 444-466.

Zipf, G. K. (1949). Human behavior and the principle of least effort, Addison-Wesley Press, Oxford, United Kingdom.

Zou, S. (2013). Towards Success in TEM8, Shanghai Foreign Language Education Press, Shanghai, China.

Does this article screened for similarity?
Yes.

Conflict of interest
The Author has no conflicts of interest to declare that they are relevant to the content of this article.