The Cross-linguistic Variations in Dependency Distance Minimization and its Potential Explanations

Ruochen Niu
Beijing Language and
Culture University
China
niuruochen@126.com

Yaqin Wang[✉]
Guangdong University of
Foreign Studies
China
wyq322@126.com

Haitao Liu
Zhejiang University
China
htliu@163.com

Abstract

Dependency distance minimization (DDM) is a preference for short syntactic dependencies observed in natural languages. While DDM has been mainly studied as a language universal, languages also seem to exhibit considerable differences in the extent of DDM. The current study adopts quantitative methods and dependency treebanks of four distinct languages to investigate the cross-linguistic DDM variations and its possible explanations. It was found that: (i) the cross-linguistic DDM variations can be captured by a parameter in the fitting functions of the dependency distance distributions; (ii) there seems to be a trade-off relation between the syntactic complexity and morphological difficulty of a language. This attempt offers new insights into the study of DDM and points out the necessity of studying language as a multi-layer complex adaptive system.

1 Introduction

Language is a human-driven multi-layer complex adaptive system (e.g., Liu, 2018). “Human-driven” highlights the motivation of its “adaption”: while abstract rules and models can be used to describe language, language in use is never static, but constantly accommodating to the constraints of human cognition and the needs of communication. “Multi-layer” and “complex” suggest that language is composed of many subsystems (e.g., syntax and morphology), and the interaction between the subsystems makes language as whole far more complex than the sum of its components. Due to the afore-mentioned nature of human languages, it is necessary to employ quantitative methods against large-scale authentic data, instead of introspection upon artificially designed materials of limited types and quantity, in order to gain insights into the complexity of natural languages.

With respect to linguistic complexity, syntactic complexity has aroused extensive interest due to its relatively high measurability and cross-linguistic comparability (e.g., Miestamo et al., 2008). While there are other models for syntactic processing difficulty (such as the expectation-based models, e.g., Levy, 2008), the most prominent ones are memory-based (e.g., Yngve, 1960; Hawkins, 1994; Hudson, 1995; Gibson, 1998). Such a model hypothesizes that the further the distance between two syntactically related words in a sentence, the heavier the load on the comprehender’s working memory. This is related to the nature of human working memory: either because it decays with time (hence distance) (e.g., Brown, 1958; Baddeley and Hitch, 1974), or because it has limited capacity (e.g., Cowan, 2000).

One of the strongest pieces of evidence supporting the memory-based account of language processing is the common tendency for human languages to use short dependencies (i.e., syntactic relations). This cross-linguistic commonality, often referred to as *dependency distance minimization* (or DDM, in which *dependency distance* refers to the linear distance between a word and its syntactic head), is absent in artificially generated random languages. The specific manifestations of DDM include but are not limited to (e.g., Ferrer-i-Cancho, 2004; Liu, 2008; Jiang and Liu, 2015; Futrell et al., 2015; Wang and Liu, 2017; Yan and Liu, 2022): (i) the probability of dependency distance (or DD) decreases monotonically, with dependencies formed between adjacent words (DD = 1) account for around 50% of all the dependencies, and longer dependencies (e.g., with a DD > 7) a rarity; (ii) the frequency distribution of DD fits well with the Right Truncated Modified Zipf-Alekseev model; (iii) the mean DD of natural languages do not exceed four, corresponding to the capacity limit of working memory; (iv) the afore-mentioned observations are immune to the change of intra-linguistic factors, such as sentence length, genre and annotation scheme. Altogether, these statistical findings demonstrate that DDM is a regularity of human languages at the syntactic level driven by the *Principle of Least Effort* (Zipf, 1949). As a result, DD is viewed as a valid measure of syntactic complexity, or even language processing difficulty (see Liu et al., 2017 for a review).

While research on DDM has struck the academia as language universal studies (e.g., Futrell et al., 2015), languages also seem to demonstrate considerable variations in terms of their extent of DDM. Hudson (2009) may be the first to point out this phenomenon. In a preface, Hudson (2009) stated that according to Liu (2008)'s results, the MDD of Chinese is almost twice that of English (3.662 vs. 2.543). As DD is a measure of cognitive load during language processing (e.g., Niu and Liu, 2022), this difference may have broader implications. "Why are the effects of working memory so different in the two languages?", Hudson asked. "Is it because Chinese words are easier to hold in memory, so that more words can be kept active? Or is it because Chinese speakers have less limited working memories?" He offered two potential explanations. As few efforts have been made to resolve these questions, Hudson (2017) reiterated them in a comment article, citing new supporting evidence.

While there has been few direct research that answer Hudson's questions, attempts have been made to quantify and differentiate the degree of DDM of languages using measures other than MDD. For example, Gildea and Temperley (2010) measured and compared the degree of DLM (or dependency length minimization) of English and German by employing computational methods on syntactically annotated corpora (i.e., treebanks). [1] To be concrete, several kinds of random languages were generated based on algorithms, including the optimal language that can yield the shortest possible dependency length (DL). Their results show that: (i) the actual DL of both languages are longer than the optimal language but shorter than the other random languages; (ii) for English, the actual word order bears more resemblance to the optimal language than the random language that is restricted to the principle of projectivity (percent of match: 45.4% vs. 40.5%). Whereas for German, the situation is the opposite (39.6% vs. 40%). These findings suggest that although English and German both follow the tendency of DDM, English is more optimized than German in that regard. Their conclusion is consistent with previous findings using MDD or DL as a measure of the degree of DDM (e.g., Liu, 2008; Futrell et al., 2015).

In a more recent study, Ferrer-i-Cancho et al. (2022) introduced another measure of the extent of DDM called optimality score. They achieved this by taking DDM as an optimization problem of a syntactic network formed by words (nodes) and syntactic dependencies (lines). According to their definition, random languages would have a score of zero, optimal languages (with shortest possible DD) a score of one, and maximal languages (the opposite of minimization) a negative score. Among the 93 languages they analyzed, half of the languages exhibit an optimality degree equal to or higher than 70%. Additionally, they found that languages can be arranged upon a hierarchy according to the score. Their rank largely coincides with the rank ordered by MDD.

The afore-mentioned studies provide new methods to describe, quantify and compare the cross-linguistic differences of DDM. However, they did not answer Hudson's questions directly, which is "why do languages differ in their MDD" and "what could be the underlying motivations". The manuscript therefore goes back to these fundamental questions, investigates whether there are variations among languages in terms of their DDM degrees, and explores the possible explanations by examining the other properties of the languages.

[1] Dependency length is a similar metric to MDD which is defined as the sum of all the DDs in a sentence (e.g., Futrell et al., 2015). Although DDM and DLM are preferred by researchers of different background, these two terms refer to the same thing by nature. This manuscript therefore uses DDM and DLM interchangeably as the general tendency for human languages to use short dependencies. It should be noted, however, that due to its nature, DL is more susceptible to sentence length than MDD (longer sentences almost always lead to longer DLs but not necessarily longer MDDs); this vulnerable nature makes DL a less robust measure for comparisons of syntactic complexity among sentences of different lengths (see Niu and Liu, 2022).

Based on dependency treebanks of four languages, we aim to address the two following research questions:

Question 1: Does the MDDs of the four languages exhibit considerable variations? If so, are these variations reflected in their DD distributions?

Question 2: What might be the underlying reasons for these variations? Are they related to the other properties of the languages, such as word order and morphology?

The rest of the manuscript is arranged as follows: Section 2 describes the treebanks and the measures used, Section 3 and 4 reports the findings relevant to the two research questions respectively, and Section 5 draws a conclusion of the manuscript.

2 Methodology

This manuscript adopts dependency grammar as the approach to analyze the syntactic structures of human languages (e.g., Tesnière, 1959; Mel'čuk, 1988; Hudson, 1995; Osborne and Niu, 2017). At the same time, methods of corpus linguistics and quantitative linguistics were also employed to generate an accurate and comprehensive understanding of natural language syntax. This section introduces the treebanks of the four languages, as well as the measures used to quantify the syntactic difficulty and word order preferences of the languages.

2.1 Dependency treebanks

For a study of language universals and language typology, it is important to have a diverse language sample (Croft, 2003). In other words, it is preferred to have languages that are distant in geographical distributions and genetic families. In the meantime, corpus-based studies need to consider the availability and the homogeneity (e.g., genre and annotation styles) of the syntactically-annotated corpora (i.e., treebanks). Given the above considerations, the manuscript adopted the dependency treebanks of news genre of four distinct languages, namely Chinese, Japanese, English and Czech. The information of the four treebanks is given as follows.

The Chinese treebank used in the manuscript is the Peking University Multi-view Chinese Treebank (PMT 1.0, Qiu et al., 2014). The texts of the treebank are the news of *People's Daily* from January 1st to 10th, 1998. In terms of the syntactic annotation, we chose the annotation based on dependency grammar provided by the developers to carry out the research. The Japanese treebank used in the study is the Balanced Contemporary Corpus of Written Japanese (BCCWJ, Maekawa et al., 2014) annotated in the style of the Surface-syntactic Universal Dependencies (SUD 2.7, Gerdes et al., 2018). For consistency of genre across languages, we extracted all the news texts from the corpus (marked as PN in BCCWJ) and formed our Japanese treebank. The English and Czech treebanks used for the analysis is the Prague Czech-English Dependency Treebank (PCEDT 2.0, Hajič et al., 2012). The English part of this parallel treebank is consisted of the Wall Street Portion of the Penn Treebank, and the Czech part is the sentence-to-sentence translation of the English part made by bilingual experts. Among the four layers of annotations provided by the developers, we chose the surface syntactic annotation based on dependency grammar (called a-layer) to conduct our research.

After confirming that the genre and the syntactic annotations of the four languages are consistent, we conducted a thorough trimming of the treebanks, excluding all the punctuation marks. Table 1 gives an overview of the treebanks after standardization. Note that tokens and sentences stand for the number of words and the number of sentences in a treebank, and MSL represent the mean sentence length (measured by the number of words in an average sentence in the treebank).

Language	**Tokens**	**Sentences**	**MSL**
Chinese	283,753	14,463	19.63
Japanese	308,456	16,027	19.25
English	927,027	43,920	21.10
Czech	784,419	43,920	17.91

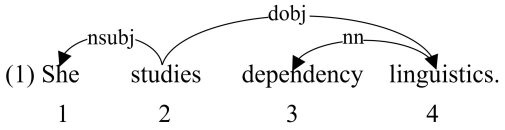
Table 1: Information about the Treebanks.

As shown in Table 1, all the treebanks are of similar magnitude, i.e., around several hundred thousand words or dozens of thousands of sentences. As for MSL (mean sentence length), Japanese and Chinese have similar and medium MSL; English has the largest and Czech the smallest MSL, although they are from a parallel treebank. This difference in sentence length may be because English and Czech have different strategies for the meaning-form correspondence, e.g., Czech adopts affixes to express meanings that can only expressed by words in English.

2.2 Dependency measures

The manuscript adopts two quantitative measures proposed within the theoretical framework of dependency grammar to describe the syntactic structures of languages. The two metrics are *dependency distance* (DD) and *dependency direction* (DDir).

Dependency grammar views syntax as direct links (called *dependencies*) between words within a sentence. To illustrate, Sentence (1) is given next:



There are four words and three dependencies in Sentence (1). Between these four words, *She* and *studies* form a subject relation, *studies* and *linguistics* a direct object relation, and *dependency* and *linguistics* a noun modifier relation. These words are linked by the arcs representing dependencies, with the arc pointing from the word that dominates (called the *head* or *governor*) to the word that subordinates (called the *dependent* or the *subordinate*).

Dependency distance and dependency direction measure the linear distance and relative position between two words forming a dependency, e.g., *She* and *studies*, respectively. The methods we used to quantify DD and DDir is based on Liu et al. (2009a). According to this method, DD equals to the linear order difference between a word and its governor. Therefore, if we define the word order of the first word of a sentence e.g., *She* in Sentence (1), as one, then *studies*, *dependency* and *linguistics* would have a word order of 2, 3, and 4, respectively. In the meantime, dependency direction is a dichotomous variable: it can be governor-final (when DD is a positive value) or governor-initial (when DD is a negative value), depending on the relative placement of the word as opposed to its governor.

Table 2 illustrates how DD and DDir are measured for a sentence. Note that sentence root (often a finite verb) is excluded from the calculations of DD and DDir because it is the only word that does not have a governor in the sentence. Based on the above methods, we can perform various calculations on all kinds of language samples, such as a sentence and a treebank.

Word	**Gov**	**DD**	**DDir**
She	studies	1	governor-final
studies	root		
dependency	linguistics	1	governor-final
linguistics	studies	-2	governor-initial

Table 2: Dependency Distance and Dependency Direction of Sentence (1).

As mentioned in Introduction, dependency distance is an index of syntactic difficulty (e.g., Hudson, 1995; Ferrer-i-Cancho, 2004; Liu, 2008; Niu and Liu, 2022; see Liu et al., 2017 for a review). Note that it is always the absolute value of DD that is used when referring to difficulty or complexity. Dependency direction, on the other hand, is a measure of word order typology, whose distribution can be used to categorize languages (e.g., Liu, 2010; Liu and Xu, 2012).

3 DDM variations of the four languages

The aim of this section is to determine whether the four languages of different language families and separate geographical locations exhibit difference in terms of their degree of DDM. The methods used are mean dependency distance (MDD), the proportion of adjacent dependencies, the distribution of DD and its fitting results.

3.1 DD-related measures

Based on the treebanks and methods introduced in above, we calculated the MDD (the arithmetic mean of all the DDs of a given language) and the proportion of adjacent dependencies of the four languages.[2] These results are compared with the findings of Liu (2008), and illustrated in Figure 1.

In Figure 1, the bar charts represent the proportion of adjacent dependencies corresponding to the left *y* axis, and the dots stand for the MDD of the language which corresponds to the right *y* axis, respectively.

[2] The proportion (or the percentage) of the adjacent dependencies is a measure of syntactic complexity related to DD. It equals to the number of dependencies formed between adjacent words (hence DD = 1) divided by the number of all the dependencies in a given sample. Given its definition, the higher this index is, the greater the syntactic complexity of the language sample.

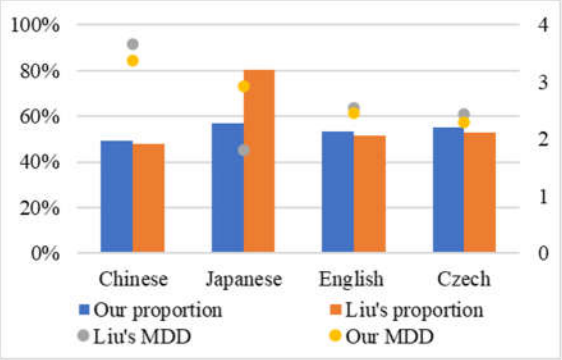


Figure 1: MDD and Proportion of Adjacent Dependencies of the Four Languages.

Figure 1 demonstrates that our results for Chinese, English and Czech are largely consistent with Liu (2008)'s based on different corpora. The only exception is Japanese: the MDD of our Japanese treebank is significantly larger than Liu (2008)'s (2.92 vs. 1.81), whereas the percentage of adjacent dependencies of our treebank is significantly higher than Liu (2008)'s (56.9% vs. 80.2%).

Given the fact that our and Liu (2008)'s treebanks for Japanese are both dependency-annotated, we believe the difference is mainly caused by genre and sentence length. The Japanese treebank used in Liu (2008) is based on dialogues, with a mean sentence length of 7.9. Compared to written texts (e.g., news used in our study), spoken texts (e.g., dialogues) enjoys higher timeliness and dynamism. To ensure that the dialogue goes smoothly, speakers may prefer to use shorter sentences and simpler syntactic structures to reduce the load on working memory during language production and comprehension (e.g., Biber et al., 1999; Wang and Liu, 2017). This stronger preference can lead to the smaller MDD and higher proportion of adjacent dependencies in Liu (2008)'s Japanese treebank.

Overall, our treebanks of the same news genre form a standard basis for cross-linguistic comparisons. Thus, the results in this subsection suggest that although the MDDs of the languages do not exceed the span of working memory (i.e., 4), they vary greatly. Among the four languages we analyzed, Chinese has the largest MDD (3.38), Japanese the second largest (2.92), English the third (2.45), and Czech has the smallest MDD (2.30). The rank of their proportion of adjacent dependencies is the opposite of their rank of MDD. These findings corroborate the MDD variations across languages. In the meantime, the MDD variations reflect that languages may enjoy different DDM degrees, with Czech being the highest, and Chinese being the lowest. The next subsection explores this phenomenon further using quantitative methods.

3.2 DD distributions and fitting results

From a mathematical point of view, the best and easiest way to understand a mean value may be to analyze its overall distribution, especially when the distribution itself has a meaning. Previous studies have found that the DD distributions of natural languages exhibit regularities (Ferrer-i-Cancho, 2004; Liu, 2007; Lu and Liu, 2016): with the increase of DD, the frequency of dependencies at a given DD decreases gradually; in the meantime, the distribution can be fitted by power functions and exponential functions.

As the size of the four treebanks are not identical (which may cause variations in the frequencies), we transformed the original frequencies into probabilities,[3] but the fitting was based on the original frequency data. Figure 2 illustrates the probability distributions of DD of the four treebanks (in which the *y* axis has been put on a common logarithm for a clearer contrast).

As shown in Figure 2, the probability of dependencies exhibits a decreasing trend with the increase of DD for all the four languages. However, if we look closely, then the rate of descent is not the same for different languages.

Among the four languages, the rate of decline

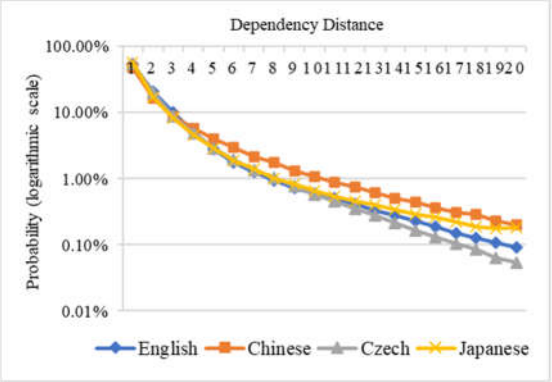


Figure 2: Probability Distribution of DD of the Four Languages.

[3] The frequency of dependencies at a given DD is the number of dependencies with a DD of that value. The probability of dependencies at a given DD equals to the frequency of dependencies at that DD divided by the frequency of all the dependencies in that language sample (e.g., a treebank, with all the possible values of DD).

in Czech is the sharpest, followed by English, Japanese and Chinese, respectively. This variation in the decline rate is further supported by statistical results: among the four languages, Czech enjoys the highest proportion of short dependencies (89.2%, with a DD ≤ 4) and the lowest proportion of long dependencies (4.52%, with a DD > 7); in contrast, Chinese has the lowest proportion of short dependencies (80.4%) and the highest proportion of long dependencies (10.4%). In other words, the results of DD distributions suggest that languages do differ in their extent of DDM. The order is Czech, English, Japanese, Chinese, from the highest to the lowest DDM extent.

Given that the results of DDM suggested by the rate of decline and MDD are the same for the four languages, we believe these variables are related. To be precise, their relations should be: the higher the rate of decline in the DD distribution, the shorter the MDD, and the higher the degree of DDM for a given language.

To further quantify the phenomenon observed above, we fitted the original frequency data of Figure 2 to the models proposed in the literature (e.g., Lu and Liu, 2016). Tables 3 and 4 present the fitting results to an exponential function and a power function, respectively:

In Tables 3 and 4, *a* and *b* are the parameters of the fitting function, and R^2 the coefficient of determination that indicates the goodness of fitting.[4] By close examination, we found that parameters *a* and *b* in both functions are a positive and negative value, respectively, which fluctuate within a limited range. As the DD distributions shown in Figure 2 exhibits a decreasing trend, we speculate that parameter *b* which is negative can reflect the rate of decline in the DD distributions and thus predict the degree of DDM of different languages.

Language	***a***	***b***	***R*²**
Chinese	0.18	-0.25	0.93
Japanese	0.14	-0.26	0.88
English	0.18	-0.29	0.92
Czech	0.21	-0.32	0.95

Table 3: Fitting the DD Distribution to the Exponential Function $y = ae^{bx}$.

Language	***a***	***b***	***R*²**
Chinese	0.67	-1.85	0.99
Japanese	0.66	-1.99	0.99
English	0.88	-2.20	0.99
Czech	1.09	-2.37	0.98

Table 4: Fitting the DD Distribution to the Power Function $y = ax^b$.

The results in Tables 3 and 4 support our speculation: the rank of parameter *b* in both fitting functions is Chinese, Japanese, English and Czech, from the largest to the smallest, which coincides with the rank of their MDD, and just the opposite of the rate of decline in their DD distributions. It should also be noted that while the value of parameter *b* varies among languages, its variation is within a very limited range. This may be because the limitations of human working memories have kept the preference for short dependencies quite consistent in languages.

In Section 3, we found that the syntactic difficulty (measured by MDD and proportion of adjacent dependencies) of the four languages exhibit considerable variations. By taking a close look at their DD distributions, we revealed the relation between the decline rate in the DD distributions, the MDD and the DDM extent (or degree) of a language. It was found that a parameter in the fitting function to the DD distributions can be used to describe and predict the DDM degree of languages. Precisely, the smaller this parameter, the higher the decline rate and DDM degree, but the lower the syntactic difficulty of a given language (corresponding to smaller MDD and higher proportion of adjacent dependencies). As a whole, this method based on the DD distributions provides a complement to the measures of DDM degree proposed in previous research (e.g., Gildea and Temperley, 2010; Ferrer-i-Cancho et al. 2022).

4 Potential explanations

The aim of this section is to explore the potential explanations of the cross-linguistic DDM variations from other properties of the languages. That "language is a system" has been well acknowledged since Saussure (1959). In the recent dec-

[4] It is generally believed that the fitting is good when $R^2 >$ 0.7; in the meantime, the greater the value of R^2, the better the model fits the data (e.g., Köhler et al., 2005).

ade, linguists have come to realize that language is not only a system, but a dynamic and multi-layer complex adaptive system (e.g., Liu, 2018). This suggests that in addition to syntax that has been the focus of the current study, other subsystems of a language, e.g., phonology and morphology, are also subjected to the influence of the biological, psychological, social and even cultural factors of human beings.

As a result, exploring the potential explanations of the cross-linguistic DDM differences from properties other than syntax is not only conducive to the understanding of the relation between language and its external influencing factors, but also helpful for understanding how different subsystems interact with each other in language comprehension and production. As a tentative attempt, this manuscript looks into the properties of the four languages in terms of word order and morphology. These two features are part and parcel of linguistic typology (e.g., Greenberg, 1963; Comrie, 1989; Croft, 2003).

4.1 Word order properties

To begin with, dependency direction and MDD distributions of the four languages were calculated and presented in Figure 3. Note that the bar charts illustrate the proportion of a word order (e.g., governor-final dependencies) corresponding to the left *y* axis,[5] and the dots represent the MDD of a given word order which corresponds to the right *y* axis, respectively.

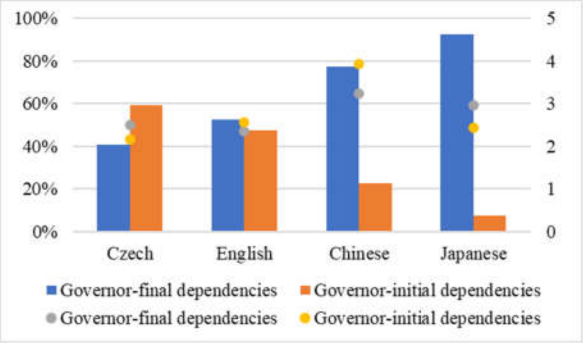


Figure 3: Dependency Direction and MDD Distribution of the Four Languages.

Figure 3 shows that the word order (or dependency direction) distributions of the four languages are quite different. Along the direction of the *x* axis (from left to right), the proportion of governor-final dependencies exhibits a growing trend, whereas the proportion of governor-initial dependencies a decreasing trend. In other words, the results indicate that from the perspective of word order typology, the relation between Czech and Japanese are the most distant among the four languages. This is consistent with Liu (2010).

Regarding MDD and head-finality, Futrell et al. (2020) proposed that languages with a higher portion of head-final dependencies tend to have larger MDD in general. This hypothesis is not supported by our results, however, because according to their prediction, the rank of MDD from the largest to the smallest would be Japanese, Chinese, English and Czech, which is inconsistent with the MDD results presented in Figure 1 and Liu (2008).

With respect to the relation between word order and the MDD of that word order, Yadav et al. (2020) found that dependencies in the word order consistent with the default word order of the language tend to have greater complexity (i.e., larger MDD). Note that the default word order refers to the ternary relation between subject, verb and object, e.g., SVO or SOV. In their study, it was found that for Czech and English, the MDD of the governor-initial dependencies is larger (as compared to that of the governor-final dependencies), whereas the MDD of the head-final dependencies is larger for Japanese. This is because Czech and English have the SVO word order and Japanese the SOV word order. Yadav et al. did not study Chinese, but according to their hypothesis, Chinese would have a larger governor-initial MDD, due to its SVO word order.

However, our results illustrated in Figure 3 do not support Yadav et al.'s findings and hypothesis. The problem is mainly surrounded at Czech: while Yadav et al. show that Czech has a larger MDD of the head-initial dependencies, our results and Liu et al. (2009b)'s results based on different treebanks both demonstrate the opposite. By analyzing the origin and the annotation styles of the treebanks, we assume that this inconsistency is mainly caused by the UD-style treebanks used by Yadav et al. (2020) (see Osborne and Gerdes, 2019 for a critique of this annotation style).

[5] The proportion of governor-final (or governor-initial) dependencies of a given language equals to the frequency of governor-final (or governor -initial) dependencies divided by the frequency of all the dependencies (the sum of governor-final and governor-initial dependencies).

This subsection examines two hypotheses proposed in previous studies regarding word order and MDD. However, the results of our study did not support them. From the other perspective, the findings suggest that the DDM variations of the four languages observed in Section 3 could not be explained by their word order properties.

4.2 Morphological properties

Many studies related to DDM use the morphological properties of languages as an explanatory factor, especially in those that have found anti-locality effects which seems to challenge the memory-based accounts (e.g., Gildea and Temperley, 2010; Futrell et al., 2020; Ferrer-i-Cancho et al., 2022). However, few studies have investigated the relation between the morphological complexity of a language and its degree of DDM directly. The aim of this subsection is therefore to fill the gap.

To investigate whether the extent of DDM of a language can be explained by its morphological properties, we adopted a framework to categorize the languages according to their word formation features (Aikhenvald 2007: 8). This framework has two parameters: one is the number of morphemes in a word, which classifies languages into analytic, synthetic, and polysynthetic ones; the other is the combining techniques of the morphemes within the word. This second parameter distinguishes between isolating, agglutinating and fusional languages.

If we use a coordinate system to represent the framework, with each axis standing for one parameter, then the morphological features of the four languages can be illustrated using Figure 4.[6]

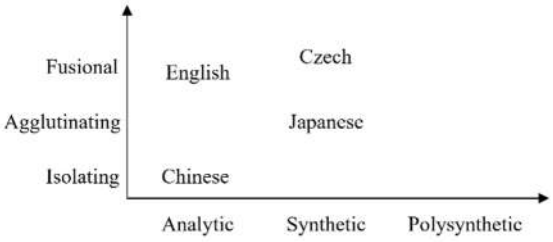


Figure 4: Morphological Classification of the Four Languages.

Note that the x and y axes stand for the classification of languages according to the number of morphemes per word and the techniques of joining morphemes within the word, respectively. Along the direction of the x axis (from left to right), the number of morphemes per word increases, hence greater morphological complexity; along the direction of the y axis (from bottom to top), the transparency of morphological boundaries is less clear, hence greater effort is needed to process and acquire word forms.

Therefore, if we view the direction of each axis is as greater complexity and the complexity caused by the joining techniques of morphemes (y axis) greater than that induced by the number of morphemes per word (x axis) (Greenberg, 1954), then the rank of the morphological difficulty of the four languages is Chinese, Japanese, English and Czech, from the lowest to the highest. This order is just the opposite of the order by the MDD of the languages. Does this finding suggest that languages enjoy lower degree of DDM (thus greater syntactic complexity) would be compensated by lower morphological complexity? This is left for further study. But at present, it is at least safe to conclude that languages have evolved different grammatical strategies for expressing meaning and the different subsystems of a language always work together to fulfill the needs of expression and communication of its users.

5 Conclusion

Based on the dependency treebanks of the news genre of four languages, the manuscript investigates the cross-linguistic variations in dependency distance minimization (DDM) and its potential explanations. Results show that: (i) the MDD differences among languages can be described and predicted by a parameter in the fitting results of the frequency distributions of DD. To be precise, the smaller the parameter b in the power law and exponential functions, the higher the rate of decline of the DD distributions, the lower the syntactic difficulty (i.e., smaller MDD and lower proportion of adjacent dependencies), and the higher the degree of DDM. This finding provides new methods and insights for the measurement of DDM extent across languages; (ii) among the four languages, Chinese has the largest MDD, followed by Japanese and English, and Czech the smallest MDD. This order is consistent with the

[6] The classification in Figure 4 is aimed to illustrate a gradient degree rather than an absolute category. While linguists may have different opinions on the extent of fusion (or inflection) of English, especially when compared to languages with much more inflectional changes like Czech; most of them agree that English is different from languages that do not have inflectional changes at all like Chinese (Haspelmath et al., 2005).

results of previous studies using different corpora; (iii) the DDM variations among languages cannot be explained by their word order features, but may be explained by their morphological features. Concretely, languages with a higher syntactic complexity (e.g., Chinese) seems to have a lower morphological complexity and vice versa. Further validation is needed to draw a conclusion, though.

To summarize, the findings and discussions of this attempt not only are conducive to understanding the commonalities and peculiarities of natural language syntax, but also point out the necessity of studying language as a multi-layer complex adaptive system. However, it should be noted at the same time that the current manuscript still has some limitations. Further studies should employ a more standard measure (e.g., syllables) to quantify the morphological complexity of languages against more language types.

Acknowledgements

This research is partly supported by the Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (22YBB28), and partly supported by the Social Science Foundation of the Education Ministry of China (Grant No. 21YJC740060), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (22JJD740018), and the Philosophy and Social Science Planning Project of Guang-dong Province (Grant No. GD23YWY02).

References

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. In T. Shopen (Ed.), *Language Typology and Syntactic Description* (Vol. 3, pp. 1–65). Cambridge: Cambridge University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology* (Second edition). Chicago: The University of Chicago Press (First edition published in 1981).

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Croft, W. (2003). *Typology and universals* (Second edition). Cambridge: Cambridge University Press (First edition published in 1990).

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*(5), 056135.

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., & Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, *105*(1), 014308.

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences (PNAS)*, *112*(33), 10336–10341.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 66–74.

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

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310.

Greenberg, J. H. (1954). A quantitative approach to the morphological typology of language. In R. F. Spencer (Ed.), *Method and perspective in anthropology: Papers in honor of Wilson D. Wallis* (pp. 192–220). Minneapolis: University of Minnesota Press.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language* (pp. 58–89). Cambridge, MA: The MIT Press.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., … others. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. *Proceedings of the 8th International Conference* *on Language Resources and Evaluation (LREC 2012)*, 3153–3160.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of language structures*. Oxford: Oxford University Press.

Hawkins, J. A. (1994). *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

Hudson, R. (1995). *Measuring syntactic difficulty*. Retrieved from http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf (Accessed 12/15/2018)

Hudson, R. (2009). Foreword. In L. Haitao, *Yīcún yǚfă de lĭlùn yǚ shíjiàn [Dependency grammar: From theory to practice]* (pp. vii–x). Beijing: Science Press.

Hudson, R. (2017). Cross-language diversity, head-direction and grammars. Comment on “Dependency distance: A new perspective on syntactic patterns in natural languages” by Haitao Liu et al. *Physics of Life Reviews*, *21*, 204–206.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency treebank. *Language Sciences*, *50*, 93–104.

Köhler, R., Altmann, G., & Piotrowski., R. G. (Eds.). (2005). *Quantitative Linguistik. Ein internationales Handbuch [Quantitative linguistics. An international handbook]*. Berlin: Walter de Gruyter.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics*, *15*, 1–12.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, *120*(6), 1567–1578.

Liu, H. (2018). Language as a human-driven complex adaptive system. *Physics of Life Reviews*, *26–27*, 149–151.

Liu, H., Hudson, R., & Feng, Z. (2009a). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*, *5*(2), 161–174.

Liu, H., & Xu, C. (2012). Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics*, *48*(4), 597–625.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.

Liu, H., Zhao, Y., & Li, W. (2009b). Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics*, *45*(4), 495–509.

Lu, Q., & Liu, H. (2016). Yīcún jùlí fēnbù yoǔ guīlǜ ma [Does dependency distance distribute regularly]? *Journal of Zhejiang University (Humanities and Social Sciences)*, *46*(4), 63–76.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, *48*(2), 345–371.

Mel’čuk, I. A. (1988). *Dependency syntax: Theory and practice*. Albany: State University of New York Press.

Miestamo, M., Sinnemäki, K., & Karlsson, F. (Eds.). (2008). *Language complexity: Typology, contact, change*. Amsterdam: John Benjamins Publishing.

Niu, R., & Liu, H. (2022). Effects of Syntactic Distance and Word Order on Language Processing: An Investigation Based on a Psycholinguistic Treebank of English. *Journal of Psycholinguistic Research*, *51*(5), 1043–1062.

Osborne, T., & Gerdes, K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics*, *4*(1), 1–28.

Osborne, T., & Niu, R. (2017). The Component Unit. Introducing a Novel Unit of Syntactic Analysis. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 165–175.

Qiu, L., Zhang, Y., Jin, P., & Wang, H. (2014). Multi-view Chinese treebanking. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 257–268.

Saussure, F. de. (1959). *Course in general linguistics* (C. Bally & A. Sechehaye, Eds.; B. Wade, Trans.). New York: The Philosophical Library, Inc.

Tesnière, L. (1959). *Elements of structural syntax*. Paris: Klincksieck.

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, *59*, 135–147.

Yadav, H., Vaidya, A., Shukla, V., & Husain, S. (2020). Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science*, *44*(4), e12822.

Yan, J., & Liu, H. (2022). Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica*, *76*(2), 406–428.

Yngve, V. H. (1960). A model and a 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: An introduction to human ecology*. Cambridge: Addison-Wesley Press, Inc.