How syntactic analysis influences the calculation of mean dependency distance: Evidence from the enhanced dependency representation

Tsy Yih
Department of Linguistics
Zhejiang University
yezi_leafy@hotmail.com

Jianwei Yan
Department of Linguistics
Zhejiang University
yanjianwei@aliyun.com

Haitao Liu✉
Department of Linguistics
Zhejiang University
lhtzju@gmail.com

Abstract

Based on Yan & Liu (2022), the present paper further explores how syntactic analysis, or the annotation scheme of dependency treebanks, affects mean dependency distance (MDD). By comparing the treebanks of 16 languages with both basic (BUD) and enhanced universal dependency (EUD) representations, we find that the MDD measured by the EUD representation is statistically larger than that of BUD. The main distinction between the two representations lies in the treatment of three constructions: coordinate structures, relative clauses, and pivotal constructions. However, a closer look at the data reveals that these three analyses in EUD do not necessarily have longer MDD in single sentences: The enhanced analysis of coordinate structures and pivotal constructions statistically have a significant contribution to the increase of MDD, while that of relative clauses contributes the least. We conclude with the factors that may affect the change of MDD, including both internal, structural ones and external ones, such as the context of the text (in the form of stochastically intervening dependents) and the language type (in the form of word order type).

1 Introduction

Mean Dependency Distance (MDD), defined as the sum of all dependency distances divided by the number of dependency relations, is a measure based on the dependency structures of sentences. It has received much attention in the last two decades (Liu, 2008; Futrell et al., 2015; Jiang & Liu, 2015). Previous studies have shown a general tendency for natural languages to possess statistically smaller MDD than languages generated by a number of random baselines (Liu, 2008; Futrell et al., 2020). Therefore, it is generally considered to be a metric to reflect syntactic complexity and the limit of human memory, and many studies have attempted to explain its inner mechanism (Temperley, 2008; Gildea & Temperley, 2010; Liu et al., 2017), such as being the result of the Principle of Least Efforts (Zipf, 1949). MDD is known to be subject to many factors, such as language type (Liu, 2008), sentence length (Jiang & Liu, 2015), chunking (Lu et al., 2016), genre (Wang & Liu, 2017), annotation scheme (Yan & Liu, 2022), etc.

Among all these factors, the annotation scheme differs from others in that it influences the observed value of the MDD, rather than the real value of the variable itself. An analogy is to measure the temperature with different scales, and

one would have different values. For instance, a certain temperature might be measured to have the value 40 under degree Celsius, and show the value of 104 under degree Fahrenheit. Turning back to the linguistic issue here, an annotation scheme reflects the choice of syntactic analysis. Since there are various versions of syntactic structural analysis in the linguistic literature, it is thus also important to pay attention to such effect. As we all know now, an underlying formula exists for the abovementioned case of temperature scales: (C × 9/5) + 32 = F. Likewise, in studying MDDs under different annotation schemes, one aim is also to find such relationship, although as we will show below, it is not that easy to have a function relation for the linguistic case.

Previously, Yan & Liu (2022) have made systematic investigations into how the syntactic annotation scheme affects the calculation of dependency distance by comparing UD and Surface-Syntactic Universal Dependencies (SUD), and they found that the MDDs in SUD are statistically shorter than those in UD. In their study, the four major constructions or controversial pairs where the dependency structures are different in two annotation schemes are the adposition-noun, auxiliary-verb, copula-noun/adjective, subordinator-verb pairs. In general, UD takes a content-head approach and SUD a function-head approach. For instance, in UD an adposition is the dependent of the head noun, while in SUD it is the head of the noun and serves as the linker between the verb and the noun. Hence, if the human language generally follows the Principle of Relator Being Intermediate (Dik, 1997), then the SUD analysis is bound to have shorter MDD values. From the comparison between UD and SUD have we learned that an annotation scheme can be seen as the combination of analyses of various linguistic constructions, and that the analysis of different constructions might probably have conflicting effects on MDD, making it much more complex than the one-dimensional variable of temperature. Yet, at least it would be worth accumulating more case studies in this trend at this stage.

Among all variants of annotation schemes available now, the enhanced dependencies are noteworthy. The enhanced representation was couched in de Marneffe et al. (2014) since the time of Stanford Dependencies (SD), and was later succeeded by the Universal Dependencies (UD) Initiative (Nivre et al., 2016). [1] In contrast with the basic dependencies which only allow tree structures, the enhanced representation allows graph structures or cyclic parts, and supplements additional relations.[2] Schuster and Manning (2016) later proposed a version of enhanced and enhanced++ UD representations, [3] which will be together called EUD in the present study, in contrast with the basic universal dependencies (BUD) representation.[4]

Taking the sample sentence in Table 1 as an instance, the 9th column (DEPS) of line 5 (dogs) in the enhanced format would be "2:obj|3:conj:and" rather than "3:conj", which indicates two relations. The graphic syntactic structures of the sentence in two formats are shown in Figure 1.

It would be interesting if we extend the calculation of MDD from BUD to EUD. With the additional links in the enhanced representation, the mean dependency distance of sentences and the whole treebanks, by definition, is subject to change. However, since both the number of relations and the sum of all dependency distances have changed, it is unclear whether MDDs would increase or decrease.

Hence, we put forward the following research questions:

(1) Do the enhanced MDDs increase or decrease compared with the original MDDs?

(2) What factors lead to the change of MDD in enhanced representation?

[1] The term "UD" could be ambiguous. In one sense, it refers to a specific annotation scheme, i.e., Yan & Liu's UD contrasted with SUD, or the BUD contrasted with EUD in the present study. In another sense, it stands for the whole annotation initiative (Zeman et al., 2017) following a specific format *.conllu*, which already has 202 treebanks of 114 languages till v2.8 (https://universaldependencies.org/).

[2] We simply use the term "relations" or "links" rather than "dependency relations" as it is hard to say if there is superiority between two words.

[3] For more information, the reader can also refer to https://universaldependencies.org/u/overview/enhanced-syntax.html.

[4] Punctuations are generally not included in the calculation of MDD.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
# text = I love cats and dogs.									
1	I	I	NOUN	PRP	_	2	nsubj	2:nsubj	_
2	love	love	VERB	VBP	_	0	root	0:root	_
3	cats	cat	NOUN	NNS	_	2	obj	2:obj	_
4	and	and	CCONJ	CC	_	5	cc	5:cc	_
5	dogs	dog	NOUN	NNS	_	3	conj	3:conj	_
6	.	.	PUNC	.	_	2	punct	2:punct	_

Table 1. The conllu format of a sample sentence

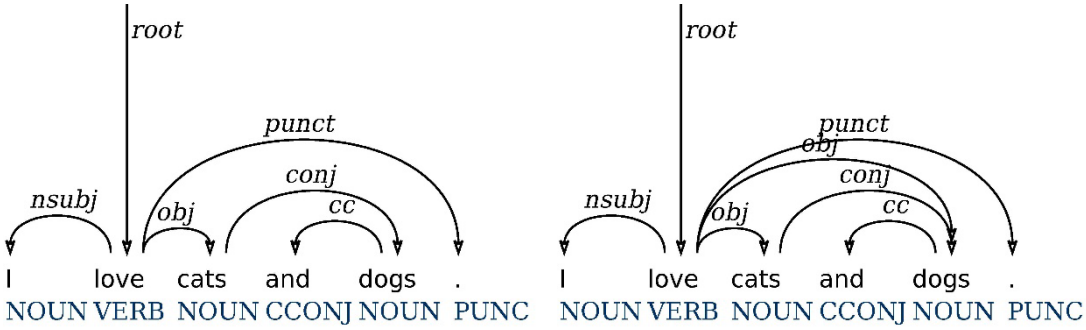


Figure 1. The basic UD annotation (the left panel) and its enhanced version (the right panel)

Language	Genera	Corpus	Genre	Sentences	$Tokens_{sp}$[5]
Arabic	Semitic	PADT.test	News	672	25911
Belarusian	IE, Slavic	HSE.test	Mixed	1020	12935
Bulgarian	IE, Slavic	BTB.test	Fiction, legal, news	1111	13451
Czech	IE, Slavic	PUD	News, wiki	984	15737
Dutch	IE, Germanic	LassySmall.test	Wiki	765	9441
English	IE, Germanic	PUD	News, wiki	993	18609
Estonian	Uralic, Finnic	EWT.test	Blog, social, web	866	10404
Finnish	Uralic, Finnic	TDT.test	Mixed	1515	17463
Italian	IE, Romance	ISDT.test	Legal, news, wiki	481	9217
Latvian	IE, Baltic	LVTB.test	Mixed	1852	23108
Lithuanian	IE, Baltic	ALKSNIS.test	Mixed	671	8774
Polish	IE, Slavic	PUD	News, wiki	1000	15731
Slovak	IE, Slavic	SNK.test	(Non-)Fiction, news	1013	10677
Swedish	IE, Germanic	PUD	News, wiki	993	17025
Tamil	Dravidian	TTB	News	600	8581
Ukrainian	IE, Slavic	IU.test	Mixed	816	13227

* The treebanks with more than three subgenres were recorded as having a mixed genre in the table.

Table 2. Basic information of the treebanks

[5] The subscript $_{sp}$ is the shorthand for *sans punctuation*, i.e., without punctuations, contrasted with $_{cp}$ – con punctuation. We devise these abbreviations for the author to report clearly and for the reader to obtain the correct information about the corpora directly. The Latin prepositions are preferred over *with* and *without* in English since they have the same initial letter "*w*".

2 Material and Methods

2.1 Material

We picked out all of the language samples that have enhanced dependencies and, most importantly, multiple dependencies from the Universal Dependencies initiative [6]. [7] As a result, 16 languages remained, as shown in Table 2. We selected the test sets of these languages to have a controllable size. [8] The only exception is Tamil's TTB: We included all of its test, training, and development sets to ensure that the size of the treebank is comparable with other treebanks.

2.2 Methods

For calculating the dependency distance of an enhanced dependency treebank, we shall start by calculating the dependency distance of a sentence. Here, Liu's (2008) approach was adopted. Formally, let $w_1...w_i...w_n$ be a word string of length $_n$. For any dependency relation between the words w_x and w_y ($x \geq 1, y \leq n$), if w_x is a head and w_y is its dependent, then the dependency distance (DD) between them is defined as the absolute value of the difference $|x - y|$. Therefore, the mean dependency distance (MDD) of a sentence is defined as:

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

where n is the number of words in a sentence and DD_i is the dependency distance of the i-th dependency relation of the sentence. Another way to define the sentential MDD is as follows:

$$MDD(\text{sentence}) = \frac{1}{m}\sum_{i=1}^{m}|DD_i| \quad (2)$$

where m is the number of all dependency relations, which is probably either equal to or larger than $n-1$. Note that there is a separate line for the root node in *conllu* format, whereas its dependency distance is zero.

Based on the second definition, the MDD of the whole treebank can be defined as:

$$MDD(\text{treebank}) = \frac{1}{M}\sum_{i=1}^{M}|DD_i| \quad (3)$$

where M is the whole relations in a treebank, which is equal to the sum of relations in each sentence. By such definition, the MDDs in both basic and enhanced representations are each a special case. In doing so, it is compatible and comparable for both cases. According this formula, the MDD of the example sentence *I love cats and dogs* in the last section in BUD representation is (1 + 1 + 2 + 1) / 4 = 1.25, while the MDD of the same sentence in EUD representation is (1 + 1 + 2 + 1 + 3) / 5 = 1.6. In this case, the MDD of EUD is higher than that of BUD.

We processed the treebank in Microsoft Excel by importing the *.conllu* format treebanks into worksheets, and did the statistical analysis by the R language[9]. The procedure of our data processing is as follows: We first deleted three kinds of sentences: The first kinds includes those which only have one root word except for punctuations, where the dependency distance cannot be calculated. The second kind contains sentences where punctuations are heads of other tokens, which causes problems when deleting punctuations. These sentences were deleted because they are hard to deal with if the language is unintelligible to us. The third kind consists of those with empty nodes because there is divergence in the treatment of the position of the empty node. For instance, the English PUD treebank duplicates the node right after the original node, while in other treebanks, the empty node can appear in any supposed position in the sentence. Yet the calculation of dependency distance relies on the exact position of words. Hence the indeterminacy of the surface position of empty nodes could be problematic. The second step was to convert all the values in the ID and HEAD columns into a relative reference in Excel for the ease of the next step. The third step was then to delete all the punctuations, a treatment

[6] All the treebanks are available from https://universaldependencies.org/.

[7] The descriptions of some treebanks claim to have enhanced dependencies, while they are simply the copy of basic dependencies without additional links.

[8] All the PUD (Parallel Universal Dependencies) treebanks only have the *test* set. Therefore they were not elucidated in the table.

[9] https://www.r-project.org/.

following Jiang & Liu (2015) and other previous studies for comparison. Since the position numbers are now relative references, they will change automatically after the rows containing punctuations were deleted. Then we calculated MDDs for both the basic and enhanced representations. Finally, the results were exported and put into R for statistical analysis if necessary.

3 Results and Discussion

3.1 Enhanced MDDs Compared with Basic MDDs

Table 3 shows the MDDs of 16 languages in both BUD and EUD annotation schemes. It can be seen that in all languages the enhanced MDDs are higher than basic MDDs, although the increments in each language are different. A paired one-sided Wilcoxon test shows that the enhanced MDD is significantly greater than the basic MDDs ($V = 0$, $p = 1.526\text{e-}05 < 0.05$).

Theoretically, the enhanced MDD is not bound to be larger than the basic MDD. When we add a new link, if the new link is an adjacent one or it has a smaller dependency distance than the original MDD of that sentence, then the whole MDD is supposed to decrease by maths. This is not a rare thing since Liu (2008), Jiang & Liu (2015) have found that adjacent relations are very common in natural language and would take up about 50% of the whole dependency relations. Futurell (2019) also argued that adjacent relation, or so-called information locality, is preferred in the structuring of language. If a new link is added by chance, then it is very likely to be adjacent. Our results, however, have revealed that the MDD goes up, indicating that the additional relations are generally long-distance ones rather than adjacent ones. To put it differently, the results seem to show that the original annotation scheme itself tends to adopt an analysis that keeps the short dependencies and omits the long-distance ones.

However, it is noteworthy that it is just statistically the EUD representations manifest longer MDDs, while there are also many single sentences with shorter MDDs, such as in (1) where the reanalysis of relative clauses in EUD plays a part.

(1) For those who follow social media transitions on Capitol Hill, this will be a little different. (English PUD)

BMDD: 2.8667

EMDD: 2.8125

A second aspect worth mentioning is that if we arrange the table in an ascending or descending order according to the MDDs before and after enhancement, the languages will be in different orders, which coincides with Yan & Liu (2022)'s finding in comparing the UD and SUD annotation scheme. This indicates that the MDD is affected by both annotation scheme and language type (e.g. head-final or head-initial). Otherwise, the orders in different representations should be the same. Hence, it is the interaction of these two factors that decide the value of MDDs. What is the nature of annotation scheme then and what part does it play in determining MDDs?

In the next section, we take a closer look at the distinction between two representations and explore what constructions have led to the increase of MDDs.

3.2 The Constructions Contributing to the Change of MDDs

As the results in the last section have indicated that the enhanced MDDs are longer than the basic MDDs in most languages, it is then natural to inquire what factors have contributed to the increase of MDDs.

Similar to Yan & Liu (2022), we decomposed two annotations schemes into constructions of which they have different analyses. The four previous phenomena do not have a distinct analysis in EUD and the cycles are not recovered in the enhanced analysis. We had a different set of constructions. Table 4 shows all types of enhancement of EUD given by Schuster and Manning (2016).

Language	Basic MDD	Enhanced MDD	Increase

Arabic	3.195	3.806	19.12%
Belarusian	2.414	2.758	14.25%
Bulgarian	2.304	2.455	6.55%
Czech	2.391	2.508	4.89%
Dutch	2.676	2.895	8.18%
English	2.528	2.715	7.40%
Estonian	2.644	2.647	0.11%
Finnish	2.307	2.633	14.13%
Italian	2.519	2.787	10.64%
Latvian	2.442	2.806	14.91%
Lithuanian	2.492	2.778	11.48%
Polish	2.226	2.415	8.49%
Slovak	2.102	2.246	6.85%
Swedish	2.473	2.647	7.04%
Tamil	2.399	2.428	1.21%
Ukrainian	2.625	3.011	14.70%

Table 3. The basic MDDs and enhanced MDDs in 16 languages

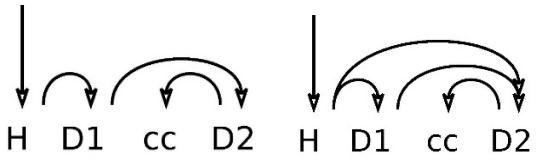
Version	Types of enhanced dependencies	Affecting MDD?
The enhanced UD representation	Augmented modifiers	No
	Augmented conjuncts	No
	Propagated governors and dependents	Yes
	Subjects of controlled verbs	Yes
The enhanced++ UD representation	Partitives and light noun constructions	Yes
	Multi-word prepositions	No
	Conjoined prepositions and prepositional phrases	Yes
	Relative pronouns	Yes

Table 4. Types of enhanced dependencies and their effects

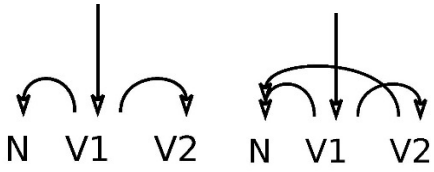
Several types of additional relations above only elaborate on the relations but do not change the dependency distances, such as those with the value of "No" in the last column. As for the rest of them, the so-called "partitives and light noun constructions" in their treatment are noted as having the dependency relation *qmod*. Yet we have not found the *qmod* relation in any annotated treebank. Besides that, the case of "conjoined prepositions and prepositional phrases" concerns empty nodes. As we have deleted the sentences with empty nodes in processing the data since they are not treated equally in different languages and might cause problems, they will not be of our concern in the present study. Therefore, the three remaining primary types of enhanced relations left are associated with coordinate structures, pivotal constructions [10] and relative clauses, which correspond to "propagated governors and dependents", "subjects of controlled verbs" and "relative pronouns", respectively, in the original terms.

Put another way, the EUD and BUD are decomposed into the combination of different analyses of these three constructions. In what follows, we will first demonstrate their treatment in two representations and then see how they affect MDDs.

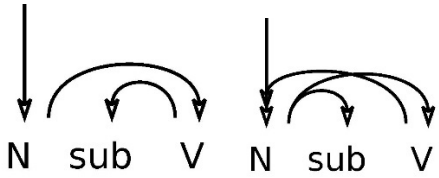
[10] The more commonly used term in the Western literature is controlled and raising structure, while we follow the use of pivotal constructions as in the Chinese linguistic literature here (Peng, 2017). In this case, the upper-level verb takes another verb as one of its syntactic argument, thereby rendering one semantic argument of the lower-level verb disappear. One has to trace its referent from the arguments of the upper-level verb. Prototypical cases include *want to*, *need to*, *start* and so forth.



(a) coordinate structures



(b) pivotal constructions



(c) relative clauses

Figure 2. The analyses of coordinate structures, pivotal constructions and relative clauses in BUD (left panel) and EUD (right panel)

As can be seen in Figure 2, in terms of the first two constructions, the enhanced structures are supergraphs of the basic structures. That is, the EUD analysis only has additional links, reflecting referential relations or functional equivalence. Hence, for the mathematical rationales presented in Section 3.1, it is easy to predict whether MDD increases or decreases based on the length of the additional links. If the length of the additional relation is longer than the original MDD in the basic setting, then it will increase the MDD in the EUD representation. However, in the case of Figure 2 (c), the relative clauses are more complicated as there are not only additional links but also changes of old links. On the one hand, there is an additional relationship between the antecedent and the root in the subordinate clause, forming a mutual dependency. On the other hand, the head of the relative pronoun is changed from the subordinate root to the antecedent. The change of DD in this local structure is |NV| + |Nsub| – |Vsub|. In the simplest case where these three elements form a continuous sequence, as |Nsub| = |Vsub| = 1, and |NV| = 2, the overall amount of increase of DD is 2. As can be seen from the graphs, the least increase of MDD in the three cases are 3, 2, and 2. Liu (2008) has shown that the MDDs in most languages fall between 2 and 3, which indicates that coordinate structures are very likely to contribute to the increase of MDDs, while the latter two constructions might probably decrease MDDs. Since the analysis above is purely theoretical and the MDD of a specific sentence may vary and is subject to the sentence length, we computed the proportions of how on earth these three constructions affect the MDD of all 16 languages dynamically, as shown in Figure 3.

In Figure 3, the black parts are those contributing to the increase of MDDs, while the white ones are those leading to the decrease of MDDs. A first sight suggests that all three constructions have the possibility to both increase and decrease MDDs, indicating that the competition between enlargement and reduction of MDD is dynamic rather than absolute.

Next, we hypothesized that the EUD analysis of the coordinate structures increases MDDs while the latter two decrease MDDs. However, the results indicate a general tendency for each construction to have an increased MDD. The overall situation does confirm that the coordinate structures have a large contribution, while relative clauses do possess less proportion. However, in many languages, the increasing part is still larger than the decreasing one (as shown by those black parts that take up more than 0.5 of all such constructions). As for the pivotal constructions, most of them increase the MDD. The one-sample sign test shows that the medians of the *conj* and *pivot* groups are greater than 0.5 (S =14, p = 0.0005 < 0.05), while that of *rel* is not significantly different from 0.5. The rank sum test shows there is no significant difference between *conj* and *pivot*, whereas both of them are greater than *rel* by one-sided tests (W = 172, p = 3.854e-06 for *conj* and *rel*, W = 167, p = 2.14e-05 for *pivot* and *rel*).

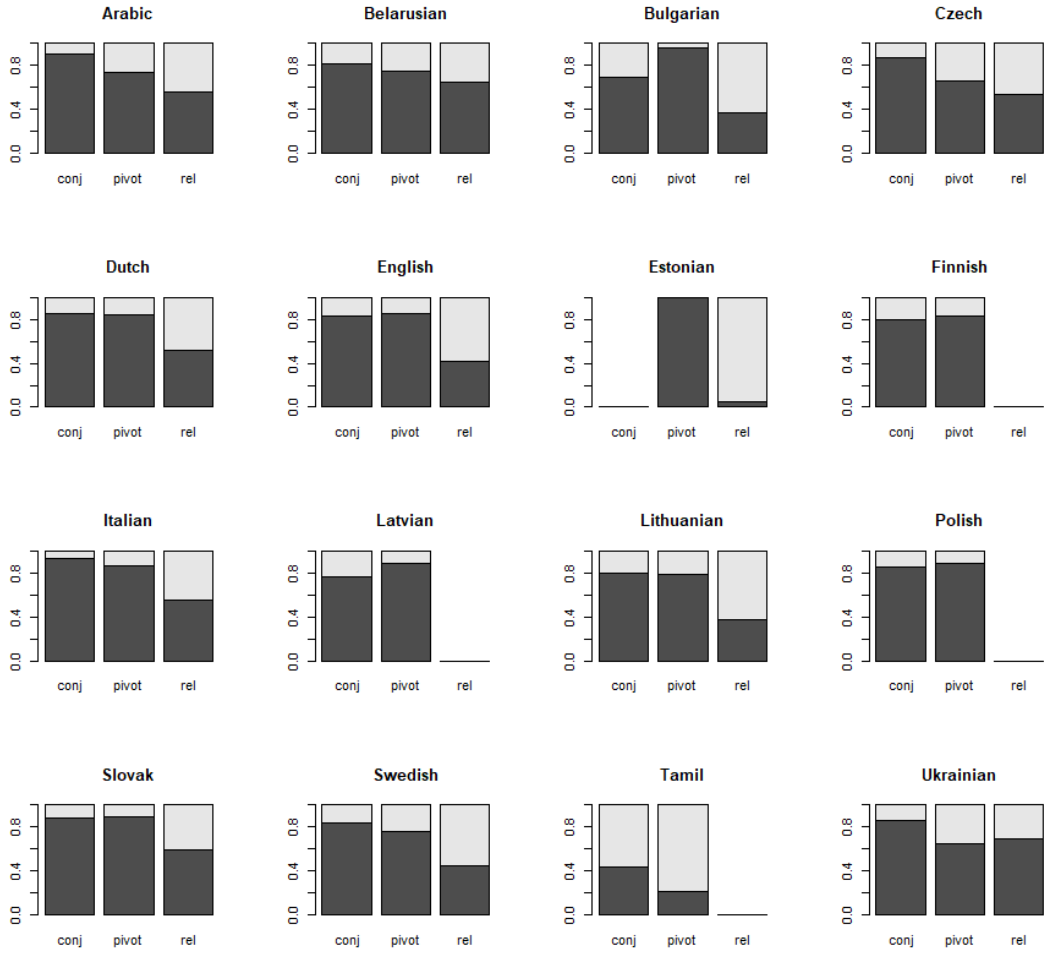


Figure 3. The proportions of the three categories in the 16 languages (black: increase; white: decrease)

Since our analyses above are based on the simplest and ideal cases, there must be other factors to be taken into consideration. The reasons can be from several aspects. From the internal, structural perspective, we have assumed a [N sub V] configuration for relative clauses where the antecedent noun, the subordinator (i.e. the relativizier or relative pronoun) and the root verb in the subordinate clause form a continuous sequence. Nevertheless, this only happens in a few restricted cases, where the verbs do not appear at the end of the subordinate clause, and the antecedents serve as the subject which is supposed to be at the beginning of the sentence. In other situations, for instance, the antecedents play the role of objects or obliques, then the words lying between them might rise dramatically. The same holds for pivotal constructions. Our analysis above is ideal and do not consider the cases such as *want to*. Even one additional particle such as *to* here would make the increase of MDD of the local structure at least to 3, which makes it probable to exceed the original MDD.

Other external factors include the content of the text and the language type. A closer look at the data reveals that there are many intervening tokens or dependents. Since the EUD representations are graph-based and have additional relations, if these links cross over a longer distance than the original MDD, they will give rise to its increase. These are not predicted from structural analysis but determined by the content that the addresser express.

Another possible factor is the language type. By language type we especially refer to the word order type. For instance, head-final languages are found

to have longer dependency distances (Futrell et al., 2020: 397). In terms of the three constructions we concern here, in those languages where the subordinate clauses are verb-final, as the antecedent will be far from the subordinate root, the EUD treatment might probably lead to an increase. As for the UD analysis of coordinate structures, the head governs the first conjunct and then the latter the second conjunct, which is related to the linear sequence. However, in a head-final language, obviously such analysis would lead to longer MDD. One might also think of an alternative annotation scheme where head-final language has a shorter MDD, such as making the last conjunct connect to the head first. Overall, we can conclude that the interaction of annotation scheme and language type would affect the values of MDDs.

There are also some problematic data. It can be found that in some languages there is no such relation at all, which indicates an annotation difference. On the one hand, there is few *conj* relation in the Estonian treebank which is also problematic, as it is almost impossible to have no coordinate construction in a not-too-small corpus. On the other hand, in the Finnish, Latvian and Polish treebanks, the EUD annotation scheme does not deal properly with relative pronouns. However, there are indeed such words as *joka* (Finnish), *kas*, *kurš* (Latvian), and *który*, *jaki* (Polish). As for the case of Tamil, it employs an affix *-a* as the relativizer, which is not suitable for the relative pronoun analysis in those European languages. This suggests that the analysis of relative clauses in UD requires reconsideration. From a cross-linguistic perspective, many languages do need relativizers, but they might not be referential as English's so-called "relative pronouns" seem to be. Therefore relativizers might also be better treated as some subordinators as those in complement clauses[11] or as a separate category, as one of UD's goals is to maximize cross-linguistic parallelism or as Croft et al. (2017) have pointed out.

[11] This same goes to the marker of adverbials clauses such as *when* and *where*. In the current version of English UD, words like *before* and *after* are treated as subordinator but *when* and *where* are treated as adverbial modifiers, which are inconsistent.

4 Conclusion

Thus far, the points to be made in the present study includes:

1. Empirically, the MDDs in the EUD representation are longer than those in the basic UD representation. Specifically, for all the three major distinctive constructions, there are cases where they increase or decrease MDD.

2. The EUD analysis of coordinate structures contributes most to the increase of MDDs, followed by that of pivotal constructions. Relative clauses, although on the whole also increase MDDs in the EUD representation, yet they have the strongest tendency to decrease among the three constructions.

3. The factors that lead to the changes include both internal, structural, and external ones, such as the content of the text (in the form of stochastically intervening dependents) and the language type (in terms of word order type). A more detailed investigation into the effects of these factors is beyond the scope of this paper and requires more comprehensive theoretical analyses and empirical validations.

To conclude, we want to re-emphasize the view that the nature of annotation scheme is the combination of analyses of various linguistic phenomena or constructions. Particularly, while the EUD representations seem to be redundant, it is simply one alternative analysis among the various possible dependency syntactic analyses. The present research is also supposed to deepen our understanding of the idea of "grammatical analysis as measurement" in language description.

A next step might be to compare more annotations schemes and decompose them into micro-parameters. Once we can manually calibrate each parameter at our will, we are likely to gain a deeper understanding of how the annotation scheme would affect the results of linguistic measurements.

Acknowledgments

This research was funded by the National Social Science Foundation in China (Grant No. 20CYY030) and the Humanities and Social Sciences Youth Foundation of the Ministry of Education of China (Grant No. 21YJC740060).

References

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 26–31, Paris.

William Croft, Dawn Nordquist, Katherine Looney, Michael Regan. 2017. Linguistic typology meets universal dependencies. In *Proceedings of The Fifteenth International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75, Bloomington, IN.

Simon C. Dik. 1997. *The Theory of Functional Grammar (2nd ed.)*. Mouton de Gruyter, Berlin.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, 112:10336–10341.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of First Workshop on Quantitative Syntax*, pages. 1–15, Paris.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length?. *Cognitive Science*, 34(2): 286–310.

Jingyang Jiang and Haitao Liu. 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.

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

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.

Qian Lu, Chunshan Xu, and Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages?. *Complexity*, 21(S2):33–41.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldbergy, Jan Hajičz, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož.

Rui Peng. 2017. *Pivotal Constructions in Chinese: Diachronic, synchronic, and constructional perspectives*. Benjamins, Amsterdam.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož.

David Temperley. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59: 135–147.

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

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA.