

# Decomposing dependency analysis: revisiting the relation between annotation scheme and structure-based textual measures

Tsy Yih  <sup>1,‡</sup>, Haitao Liu<sup>2,\*</sup>

<sup>1</sup>School of Foreign Studies, Tongji University, Shanghai, 200092, China

<sup>2</sup>College of Foreign Languages and Literature, Fudan University, Shanghai, 200433, China

\*Corresponding author. College of Foreign Languages and Literature, Fudan University, Shanghai, 200433, China. E-mail: htliu@163.com

<sup>‡</sup>Tsy Yih is the transliteration of the name of the first author in his mother tongue, Shanghai Wu Chinese. He is also known as ZI YE in Mandarin pinyin.

## Abstract

Standardized quantitative measurement of texts lies at the heart of digital approaches to humanities. Structure-based textual measures are known to be influenced by the choice of syntactic annotation schemes. Building on previous research, the present article further explores the relation between annotation schemes and the index of mean dependency distance (MDD) by comparing the treebanks of seventeen languages, respectively, within a tree representation (basic universal dependencies, BUD) and within a graphic representation (enhanced universal dependencies, EUD). Following the idea of decomposing annotation schemes into the combinations of analyses of specific constructions (coordinate structures, control constructions, and relative clauses), we design algorithms to identify them in the CoNLL-U format treebanks and explore their influences. It is found that the overall MDD of the EUD representation is statistically higher than that of BUD at corpus level, primarily affected by the coordinate structure due to its high frequency. At sentence level, all three constructions might contribute to either increased or decreased MDD, with stochastically intervening words and word order being two important determinants of the values of the measure. Finally, we propose and argue for the view that MDDs calculated under different annotation schemes should be regarded as different textual measures in nature. In sum, the present study provides another case study to deepen our understanding of the nature of syntactic annotation schemes and its relation with textual indices, which paves the way for standard measurement of texts in future humanities research.

**Keywords:** textual measures; mean dependency distance; annotation scheme; Universal Dependencies; enhanced dependencies.

## 1. Introduction

Texts occupies a central position in social sciences and humanities, especially in digital humanities (Hovy 2020, 2022; Schneider 2024). In recent years, the corpus, computational, and quantitative linguistic approaches have quantified a large number of textual measures, such as lexical richness, syntactic complexity, and so on. However, the history of scientific measurement of text is not that long as the measurement in physical or psychological sciences. Thus far, to a large extent, we still do not understand clearly how to measure texts or conduct precise and standardized measurements.

All textual measures can be divided into two types. One is independent of syntactic structures, such as the well-known type-token ratio (Johnson 1944), while the other is structure-based that takes syntactic parsing

as necessary intermediate step in the computation, including mean dependency distance (MDD) (Liu 2007), dependency direction or Liu-directionality (Liu 2010), and mean hierarchical distance (Jing and Liu 2015). In the latter case, syntactic annotation schemes play the indispensable role of scaffolds. In this article, we focus on one widely studied measure of the second type, MDD, defined as the sum of all dependency distances divided by the number of dependency relations. It is considered to reflect syntactic complexity based on the dependency structures of sentences. MDD has received much attention in the last decades in both theoretical (Hudson 1995; Liu 2007, 2008; Liu, Hudson, and Feng 2009; Liu, Xu, and Liang 2017; Ferrer-i-Cancho and Gómez-Rodríguez 2021) and applied research<sup>1</sup> (Jiang and Ouyang 2018; Ouyang and Jiang 2018; Li

and Yan 2021; Hao, Wang, and Lin 2022). To achieve higher validity and reliability in cross-study comparison and real pedagogical or clinical practice, a standard process of measuring this index is critical. Therefore, as summarized in Liu (2022: Ch. 2), a number of studies have been conducted to investigate factors influencing the values of MDD, such as sentence length (Jiang and Liu 2015), genre (Wang and Liu 2017), etc.

In addition, as is known from the literature, the choice of annotation scheme also has impacts on the calculation of the values of textual measures (Yan and Liu 2022, 2024). In early stages, since there were not as many annotation frameworks to choose from, the calculation of MDD was not very rigorous. Some previous studies (Liu 2008) employed different treebanks with rather varying standards, or even originating from different frameworks, such as those converted from phrase structures. Yet currently, several relatively standardized annotation schemes have emerged. When calculating syntactic complexity measures like MDD, it is always necessary to choose one particular framework, and the results can vary depending on the framework used. The selection of an appropriate framework and the rationale behind it warrant thorough investigation. Consequently, the influence of annotation schemes on how syntactic complexity is measured has received increasing attention.

There are a number of existent dependency frameworks. In addition to the above-mentioned UD (Nivre et al. 2016), other most famous ones include Stanford Typed Dependencies (SD, de Marneffe and Manning 2008), Surface-Syntactic Universal Dependencies (SUD, Gerdes et al. 2018), etc. From a theoretical perspective, Osborne and Gerdes (2019) argue that SD is better than UD in terms of syntactic diagnoses. Quantitatively, Yan and Liu (2019, 2022) explored how the syntactic annotation scheme affects the calculation of dependency distance by comparing UD and SUD, and found that the MDDs in SUD are statistically shorter than those in UD, in line with Liu's (2008) original foundational finding. It is noteworthy that some literature (e.g. Gerdes et al. 2018; Osborne and Gerdes 2019) held the view that the SUD annotation is preferred over UD because it better reflects the principle of Dependency Distance Minimalization (DDM), a general tendency for natural languages to possess statistically smaller MDD than random languages generated by a number of baselines (Liu 2008; Futrell, Mahowald, and Gibson 2015; Futrell, Levy, and Gibson 2020).

However, there are several limitations in previous studies. First, prior research only focuses on tree dependency structures. Yet, we are aware of no prior work that has studied MDD in graphic representation and compared the MDDs in two types of representations. As mentioned in some literature (cf Mel'čuk 1988; Hudson 2007), the

prevalence of tree syntactic structures is mainly out of computational simplicity and theoretical succinctness in the early stage of linguistic inquiry, while a more authentic picture and more informative representation should be graphic structures, involving phenomena such as mutual and multiple dependencies. Therefore, graphic representations are worth our attention. Second, although some studies (e.g. Yan and Liu 2019, 2022, 2024) have pointed out a series of distinctive phenomena in qualitative comparison between annotation schemes, it has not been investigated in detail how the specific distinct constructions between them respectively affect the calculation of MDD in quantitative manner. Last but most importantly, the argumentation that an annotation scheme with lower MDD (e.g. SD and SUD) is preferred over the one with higher MDD (e.g. UD) just because it better fits the principle of Dependency Distance Minimalization is open to debate. Since this universality was found within certain syntactic analytical framework, such an argument would lead to circularity. Therefore, a more appropriate view of the relation between the annotation scheme and MDD is called for.

To fill these research gaps, the present study aims to further explore the relation between annotation schemes and the textual measure of MDD, attempting to unveil its nature by extending the calculation of MDD in tree-structural, basic dependency representation to that in graph-structural, enhanced dependency representation.

Specifically, we aim to address the following research questions:

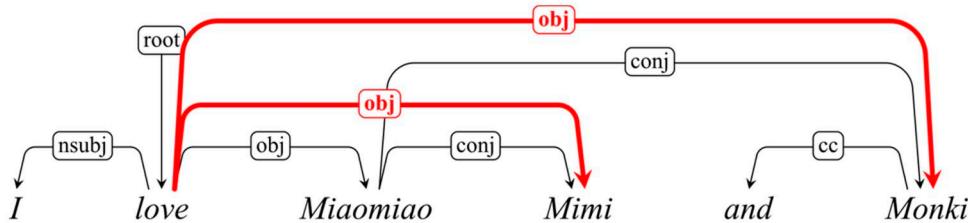
- 1) Are MDDs in the enhanced representation higher or lower than those in the basic representation?
- 2) How do distinct constructions between the two annotation schemes and other factors influence the change of MDDs in enhanced representation?
- 3) What view should we hold with regard to the relation between annotation scheme and structure-based textual measures such as MDD?

Among the three research questions, the first two are explored empirically by utilizing seventeen treebanks from the UD project, while the last one is conceptual and theoretical, which will be discussed at length. The present study intends to provide an in-depth analysis regarding the relation between annotations schemes and textual measures.

## 2. Basic and enhanced universal dependencies

### 2.1 Backgrounds

The enhanced dependency representation can be traced back to de Marneffe et al. (2014) at the time of SD,



**Figure 1.** The basic UD annotation and its enhanced version, without and with additional dependencies in red. The punctuations are omitted.

which was the annotation scheme used for the original version of well-known syntactic parser, Stanford Parser, in computational linguistics. It was then succeeded by Universal Dependencies (UD),<sup>2</sup> a treebank construction project aimed for maximal cross-linguistic applicability (Nivre *et al.* 2016). Schuster and Manning (2016) later proposed a version of enhanced and enhanced++ UD representations,<sup>3</sup> which will together be called EUD in the present study, in contrast to the basic universal dependencies (BUD) representation. The enhanced universal dependencies (EUD) representation differs from all other variants of annotation schemes available now in that it goes beyond tree structures and allows for graphic structures by adding additional dependencies. For instance, Fig. 1 shows the syntactic structures of the sentence *I love Miaomiao, Mimi and Monki* in both representations.

As can be seen, in the multiple coordinate structure *Miaomiao, Mimi and Monki*, the second (*Mimi*) and third coordinands<sup>4</sup> (*Monki*) were both originally attached to the first coordinand *Miaomiao*, as prescribed by the basic UD annotation guideline. In the enhanced representation, however, the dependency between the first coordinand and its head *love* is then propagated to the other two. Consequently, we have three obj relations in the EUD structure.

We also present it in a simplified CoNLL-U format as shown in Table 1. CoNLL-U is currently a standard data exchange format for the representation of linguistic data. In its original form, each line stands for a token, and there are ten columns each designated with a presupposed meaning. Here we only keep related columns. As can be seen, the HEAD and DEPREL columns denote the head and relation type of the dependency in basic representation, whereas the DEPS column records the enhanced dependencies, which might be single or multiple. For example, the tokens *Mimi* in line 4 and *Monki* in line 6 both have one basic dependency 3 : conj in the DEPREL column but two dependencies in the DEPS column, one simply copied from the basic dependency and another (2 : obj) propagated from the first coordinand, namely, *Miaomiao* in line 3. In addition, the dependency distance of each dependency is shown in the last column,

calculated by subtracting the token ID from its head's ID and then taking the absolute value.

We will denote the MDD calculated within the basic representation as  $MDD_{BUD}$ , and that within the enhanced representation as  $MDD_{EUD}$ . Recall that MDD is calculated as the sum of all dependency distances divided by the number of dependencies. By definition, once there is change in structure,  $MDD_{EUD}$  is bound to be different from  $MDD_{BUD}$ . 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, and what factors give rise to this change.

## 2.2 Distinctively analyzed constructions in the two annotation schemes

Yih (2023) proposed the view that ‘an annotation scheme is the combination of analyses of various linguistic phenomena’. According to this view, any annotation scheme can be decomposed into the treatment of constructions, just like a DNA or protein sequence. However, since this whole sequence would be too long to describe completely, when we want to compare just two annotation schemes, we can delimit the range to the exact constructions for which they have different analyses. For instance, in Yan and Liu’s (2022) comparison of SUD and UD, the distinctive constructions include the adposition–noun, auxiliary–verb, copula–noun/adjective, subordinator–verb pairs.

As for BUD and EUD, although there are in fact more distinctive constructions analyzed differently in two annotation schemes, in this study we only focus on the constructions that affect the values of MDDs. To this end, three major types remain, including coordinate structures, control constructions, and relative clauses, which correspond respectively to ‘propagated governors and dependents’, ‘subjects of controlled verbs’, and ‘relative pronouns’ in the original terms (Schuster and Manning 2016). In other words, after filtering, the EUD and BUD are decomposed into the combination of different analyses of these three constructions. The examples in English are given in (2).

- (2) Coordinate structures: I love cats and dogs.
- Control constructions: I like watching TV.

**Table 1.** A simplified CoNLL-U format of the sample sentence *I love Miaomiao, Mimi and Monki*, and the dependency distances of each dependencies shown in the last column.

ID	FORM	LEMMA	UPOS	HEAD	DEPREL	DEPS	DD
1	I	I	NOUN	2	nsubj	2: nsubj	1
2	love	love	VERB	0	root	0: root	-
3	Miaomiao	Miaomiao	PROPN	2	obj	2: obj	1
4	Mimi	Mimi	PROPN	3	conj	3: conj 2: obj	1 2
5	and	and	CCONJ	6	cc	6: cc	1
6	Monki	Monki	PROPN	3	conj	3: conj 2: obj	3 4

**Relative clauses:** I love cats that have white fur.

For specific structural analyses of the three constructions, see Fig. 2 for summarization. We also distinguished between several subtypes.

Figure 2 needs some specification. First, for coordinate structures or coordination, a major division is drawn between those with conjoined dependents (A1) and with conjoined dependents (A2). In the former case, an additional dependency between the head and the non-initial dependent coordinand is copied, or propagated, from the one between the head and the first coordinand, by tracking the relation of conj. In the latter case, similarly, one can ‘enhance’ the representation by supplementing a dependency between the non-initial head and the shared dependent through propagation. As for A3, coordination with propagated roots, it is in fact a special case of A2, where the propagated dependency from the supposed head is root. Note that since root is not counted in the calculation of MDD, the value of the measure would not be changed in this case. Second, in control constructions, the subject relation between the dependent and the head word of the superordinate clause is propagated to the subordinate head, which is governed by a xcomp dependency by the upper head.

As can be seen, in terms of coordination and relative clauses, the enhanced representations are supergraphs of the basic ones. That is, the EUD analysis simply adds an additional dependency and keeps the original structure unchanged, reflecting co-referential relations or propagated functional equivalence.

However, in the case of relative clauses, the situation is more complicated. The C2 subtype is supposed to be the standard analysis according to the current UD guideline. On the one hand, there is an additional, propagated dependency relation between the antecedent and the subordinate head, giving rise to mutual dependencies. On the other hand, the head of the relative pronoun is changed from the subordinate root to the antecedent with a new dependency type ref, denoting co-reference. The difference between C1 and C2 primarily lies in that the deleted original basic dependency (the dash line in C2) is kept in the C1 subtype. In fact, C1 and C2 are just alternative analyses of the

same linguistic phenomenon. As for C3, it applies to the cases where relativizers are covert, forming a circle. It should be noted that the subordinate head sometimes is not identical to the propagator, ie, the head of the propagated dependency, as shown in C4.

Despite the standard status of C2, in fact, the analysis in C1 is more consistent with the other two types of constructions, which will be shown below.

In general, the configurations of first two cases and the C1 type of relative clause can be basically unified and represented as a triad as shown in Fig. 3. A triad is a directed acyclic graph (DAG), a simplest triangle local structure in a dependency graph. It consists of three nodes and three directional edges, where each node denotes a word, and each edge a dependency relation. In a triad, here is always one node ( $W_a$ ) governing the other two, another ( $W_b$ ) being intermediate, and the rest one ( $W_c$ ) being multiply governed. There are six possibilities to linearize a three-element triad. If it is linearized in the way as in Fig. 3, the shortest path for connecting all three words is apparently  $|dep_2| + |dep_3|$ .

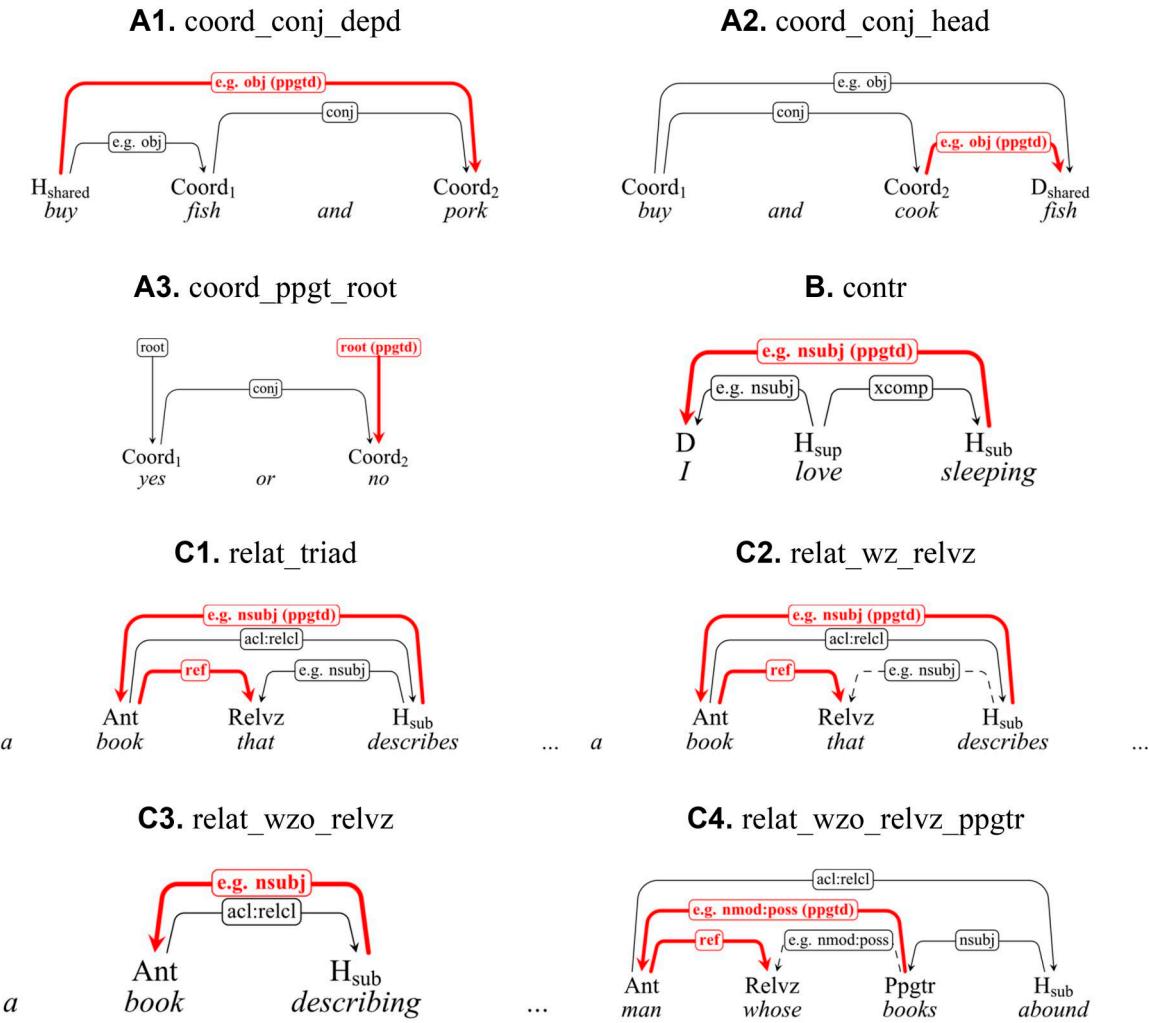
One can easily identify the triad configuration in the enhanced representation of coordinate and control structures. As for relative clauses, only C1 has a full triad, as the name of this subtype tells. Yet we can still imagine an underlying triad to process the structure of C2. With the help of the theoretical notion of triad, we developed algorithms for identifying most types and subtypes of constructions in the CoNLL-U format treebanks, which are given in Appendix 1.

After introducing the basic theoretical and structural notions regarding enhanced dependencies, we begin to address our research questions empirically in the following sections.

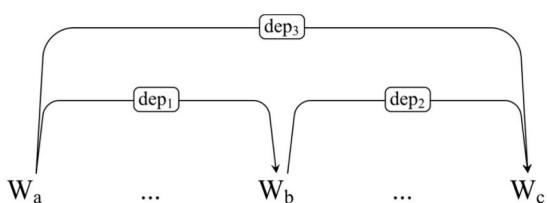
### 3. Materials and methods

#### 3.1 Material

The treebanks used in the present study were from the 2.14 version of the UD initiative (released on May 15, 2024).<sup>5</sup> After singling out all of the language samples that have enhanced dependencies and, most importantly, multiple dependencies,<sup>6</sup> seventeen languages remained. In this study, for each language, we only



**Figure 2.** Types and subtypes of three distinctive constructions between BUD and EUD. H, head; Coord, coordinand; D, dependent (coordinand); Ant, antecedent; Relvz, relativizer; <sub>sub</sub>, subordinate; <sup>sup</sup>, superordinate; ppgtd, propagated; ref, coreference. The red line denotes the additional dependency and dashed line the canceled dependency in EUD. (A1) Coordination with conjoined dependents. (A2) Coordination with conjoined heads. (A3) Coordination with propagated roots, special case of (A2). (B) Control construction. (C1) Relative clause with full triad. (C2) Relative clause with relativizer. (C3) Relative clause without relativizer. (C4) Relative clause with relativizer (subordinate head being not identical to propagator), special case of (C2).



**Figure 3.** The general configuration of a triad.

chose one treebank even if it has multiple treebanks with enhanced representation. Since the specific annotation details in different treebanks vary, mixing

heterogeneous treebanks might result in inconsistencies in the results and affect the reliability of the conclusions. The specific selection criteria were as follows: if that language has the PUD (Parallel Universal Dependencies) treebank, it would be the first priority, because PUD is a series of dedicatedly designed parallel corpora, which would be more comparable. If not, then we chose the larger treebank among all. The basic information of these corpora is shown in Table 2.

It can be seen that the sizes of treebanks are not balanced. Yet since we are not comparing different languages but how MDDs are affected by annotation schemes in each language, this would not be a problem.

**Table 2.** Background information and sizes of the treebanks.

Language	Genera	Treebank	Genre	# Original sentences #	# Remained sentences (%)	Tokens <sub>cp</sub>	Tokens <sub>sp</sub>
Arabic	Semitic	PADT	news	7,664	7,094 (92.56)	320,775	298,330
Belarusian	IE, Slavic	HSE	mixed	25,231	24,568 (97.37)	305,417	248,402
Bulgarian	IE, Slavic	BTB	fiction, legal, news	11,138	9,763 (87.65)	156,149	134,090
Czech	IE, Slavic	PUD	news, wiki	1,000	941 (94.10)	18,668	16,044
Dutch	IE, Germanic	LassySmall	Wiki	17,120	16,309 (95.26)	297,486	262,903
English	IE, Germanic	PUD	news, wiki	1,000	933 (93.30)	21,316	18,865
Estonian	Uralic, Finnic	EWT	blog, social, web	7,190	6,713 (93.37)	90,694	75,846
Finnish	Uralic, Finnic	TDT	mixed	15,136	13,394 (88.49)	202,697	173,091
Italian	IE, Romance	ISDT	legal, news, wiki	14,167	13,339 (94.16)	318,263	284,386
Latvian	IE, Baltic	LVTB	mixed	18,850	16,781 (89.02)	318,183	262,694
Lithuanian	IE, Baltic	ALKSNIS	mixed	3,642	3,274 (89.90)	70,049	57,153
Polish	IE, Slavic	PUD	news, wiki	1,000	954 (95.40)	18,433	15,779
Russian	IE, Slavic	SynTagRus	(non-)fiction, news	87,336	83,642 (95.77)	1,517,881	1,237,158
Slovak	IE, Slavic	SNK	(non-)fiction, news	10,604	10,352 (97.62)	106,203	87,073
Swedish	IE, Germanic	PUD	news, wiki	1,000	915 (91.50)	19,085	17,144
Tamil	Dravidian	TTB	news	600	597 (99.50)	10,416	9,416
Ukrainian	IE, Slavic	IU	mixed	7,092	6,401 (90.26)	123,032	99,777

1. The treebanks with more than three subgenres are noted as having a mixed genre in the table.

2. We use the notations following Yih *et al.* (2022), where the subscript <sub>sp</sub> stands for *sans punctuation* (without punctuations), and <sub>cp</sub> for *con punctuation* (with punctuation).

We preprocessed the data as follows. Four kinds of sentences in the original treebanks were removed: The first kind consists of those with empty nodes. Since in the enhanced dependencies of some tokens, the heads are empty nodes with ids such as ‘7.1’ rather than integers, these would cause problem to our calculation and were thus deleted. The second kind is those with obvious annotation errors which would affect the results. To achieve this, we employed a rule-based error checker (cf Yih and Dai 2023) to detect them, such as punctuations being used as heads of other tokens, which are impossible to exist. The third kind includes the sentences containing tokens with more than two enhanced dependencies, which would complicate the process for construction identification to a large extent in our preliminary pilot study. Fortunately, their numbers are very limited, and the influence of the deletion would be neglectable. Finally, we excluded the sentences which were supposed to contain the three phenomena under investigation, but wrong annotated and not conforming to the ideal structural analysis. In addition, the separate lines of multi-word tokens (e.g. contractions such as *you're* in English), but not the whole sentences containing them, were also removed. After several rounds of filtering as mentioned above, we kept the sentences left as our formal data source, and calculated MDDs within both the basic and enhanced representations. All statistical analyses in this research were conducted using the SciPy package<sup>7</sup> in Python. The visualizations were done either via the ggplot2<sup>8</sup> package in R,<sup>9</sup> or the Matplotlib<sup>10</sup> and Seaborn<sup>11</sup> packages in Python, considering which is more appropriate and convenient for specific cases.

### 3.2 Methods

To calculate the MDD of a specific dependency treebank, we adopted Liu's (2008) approach. Suppose there is a sentence or word string of length  $n$ ,  $w_1 \dots w_i \dots w_n$ . If there is a dependency relation between the words  $w_x$  and  $w_y$  ( $1 \leq x, y \leq n$ ), where  $w_x$  is the head and  $w_y$  is the dependent governed by  $w_x$ , then the dependency distance (DD) of this dependency is defined as the absolute value of the difference  $|x-y|$ . For simplicity, it can also be taken directly as the DD of the dependent token,  $w_y$ . Based on that, the 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<sup>12</sup> and  $DD_i$  is the dependency distance of the  $i$ th dependency relation of the sentence. The denominator is  $n-1$ , because the dependency distance of the root dependency is considered to be zero since it has no head. However, in enhanced representation, the number of dependencies is not equal to the number of words minus one. Hence, an equivalent formula to define MDD is used:

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

where  $m$  is the number of all dependency relations, which is equal to  $n-1$  in basic representation and might be larger than that in enhanced representation.

Following the second definition, the MDD of the whole treebank<sup>13</sup> thus can be extended to be defined as:

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

where  $M$  is the number of all dependencies in a treebank, which is equal to the sum of relations in all sentences. In doing so, the MDDs in both basic and enhanced representations can be captured by and considered special cases of this formula, which guarantees the compatibility and comparability.

Take the sample sentence *I love Miaomiao, Mimi and Monki* in Fig. 1 as an instance. According to formula (3), the MDD of the sentence in the BUD representation is  $(1+1+1+1+3)/5 = 1.4$ , while the MDD of the same sentence in the EUD representation is  $(1+1+1+2+1+3+4)/7 = 1.86$ . In this case, the MDD of EUD is higher than that of BUD.

To address the second research question, we also considered the DD change caused by each specific construction. To obtain that, we first identified all related examples in the treebanks of each language. Then the calculation was performed as follows:

$$DD\ change(\text{construction}) = \sum |DD_{\text{enhanced}}| - |DD_{\text{basic}}| \quad (4)$$

Still consider the sample sentence *I love Miaomiao, Mimi and Monki*. Since the tokens *Mimi* and *Monki* have enhanced dependencies, the overall DD change caused by this coordinate structure is  $DD_{\text{enhanced}}(\text{Mimi}) - DD_{\text{basic}}(\text{Mimi}) + DD_{\text{enhanced}}(\text{Monki}) - DD_{\text{basic}}(\text{Monki}) = (1+2) - 1 + (3+4) - 3 = 2 + 4 = 6$ .

## 4 Results and discussion

### 4.1 Overall MDDs

The line chart in Fig. 4 shows the overall MDDs in both basic and enhanced representations in 17 languages at corpus level. With all constructions and sentences mixed together,  $MDD_{\text{BUDs}}$  are generally higher than  $MDD_{\text{EUDs}}$ , with statistical significance confirmed by a paired one-sided Wilcoxon test ( $W = 153$ ,  $P = 7.629e-06 < 0.05$ ). Nevertheless, if we take a closer look at the sentence level, the grouped bar chart in Fig. 4 also displays that there is a large proportion of single sentences whose  $MDD_{\text{EUDs}}$  are smaller than  $MDD_{\text{BUDs}}$  (in blue) in all languages. A notable case is Tamil, in which the number of sentences with decreased MDD is larger than that with increased MDD, but the total  $MDD_{\text{EUD}}$  is still increased compared with the basic one. A more concrete English example is

given below as in (1), whose  $MDD_{\text{BUD}}$  is larger than  $MDD_{\text{EUD}}$ , in contrast to the tendency at corpus level.

(1) *For those who follow social media transitions on Capitol Hill, this will be a little different.* (English PUD, sent\_id = n01001013)

$MDD_{\text{BUD}}$ : 2.8667

$MDD_{\text{EUD}}$ : 2.8125

With the findings above in mind, in the next section, we aim to further investigate the factors that influence the change of MDD in two representations by decomposing the annotation schemes and focusing on distinct constructions.

### 4.2 Factors affecting values of MDDs at corpus level

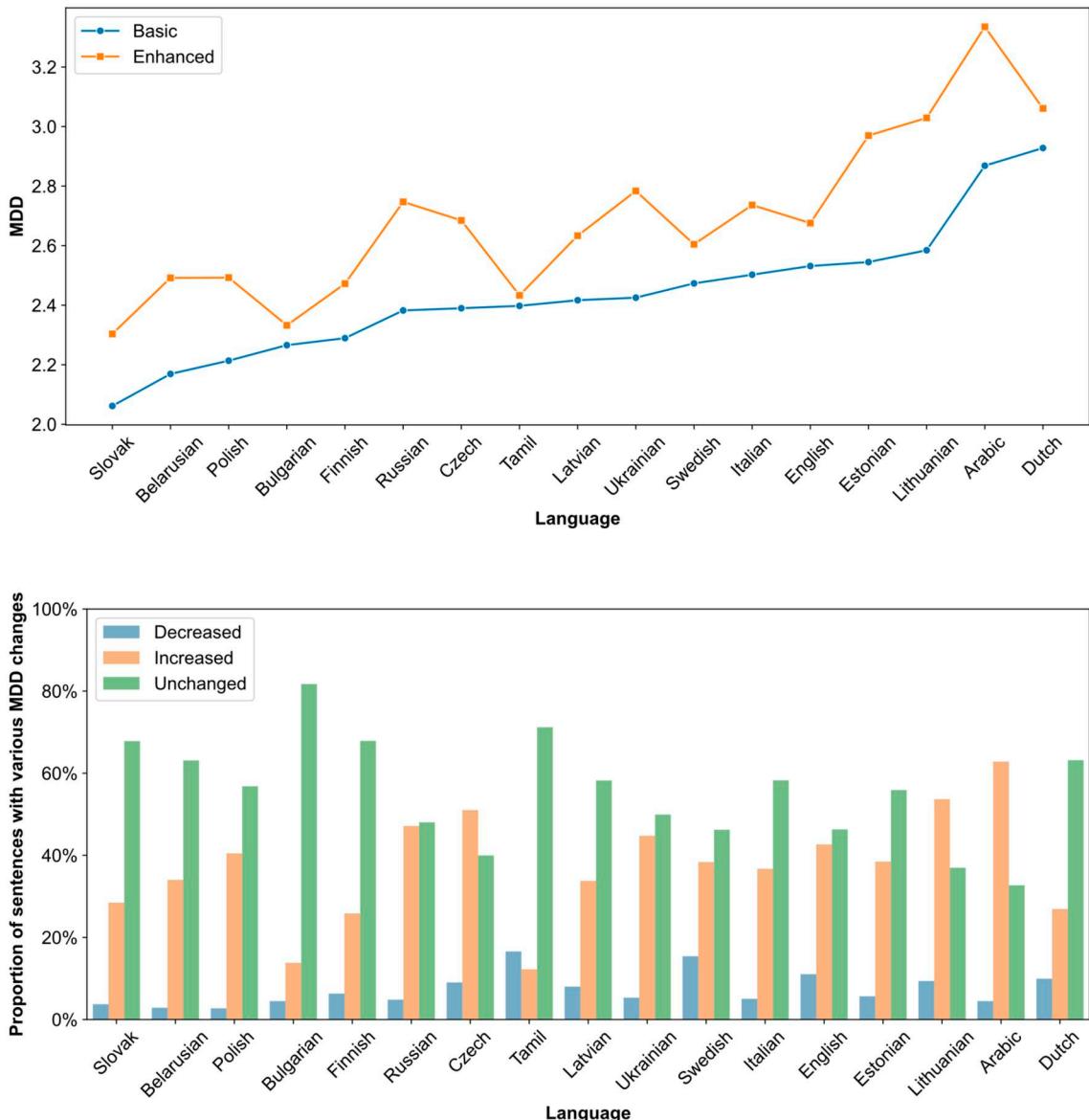
According to Yih, Yan, and Liu's (2022) conjecture, they divide factors influencing the change of MDD into internal and external ones. The former concerns the structural configuration of distinct constructions between the annotation schemes, while the latter include stochastically intervening words that lengthening the DD, and the word order type regarding specific constructions. However, in that paper they do not examine these factors quantitatively. Hence in this section, we aim to explore them in detail.

#### 4.2.1 Construction types

Table 3 lists the overall MDDs under all controlled situations of whether each construction is taken into consideration or not. A generalized linear mixed regression taking *Language* as the group variable,  $MDD \sim coord * contr * relat + (1 | Language)$ , shows that coordination is the main factor ( $P = 0.000$ ), whereas influences of control constructions and relative clauses are neglectable ( $contr: P = 0.770$ ;  $relat: P = 0.760$ ). The finding indicates that although the latter two constructions also have impact on the overall MDD, their contributions is not significant, possibly due to their low frequencies appearing in the text or low MDD changes induced by the enhanced analysis with respect to that construction.

#### 4.2.2 Relative frequencies and average DD changes of the three constructions

According to formula (3), two potential forces affecting the overall MDD are how many constructions there are and how much each of them contributes. Figure 5, thus, illustrates the relative frequencies and DD changes of the three constructions. Seen horizontally, it can be found that the coordination group is scattered on the right side of the graph, reflecting its higher relative frequencies than those of the other two groups, which is confirmed by the paired one-sided Wilcoxon signed-rank test ( $coord$  vs.  $contr: W = 105$ ,



**Figure 4.** The MDD<sub>BUDS</sub> and MDD<sub>EUDS</sub> (line chart) and the proportions of sentences with MDDs changed differently (grouped bar chart).

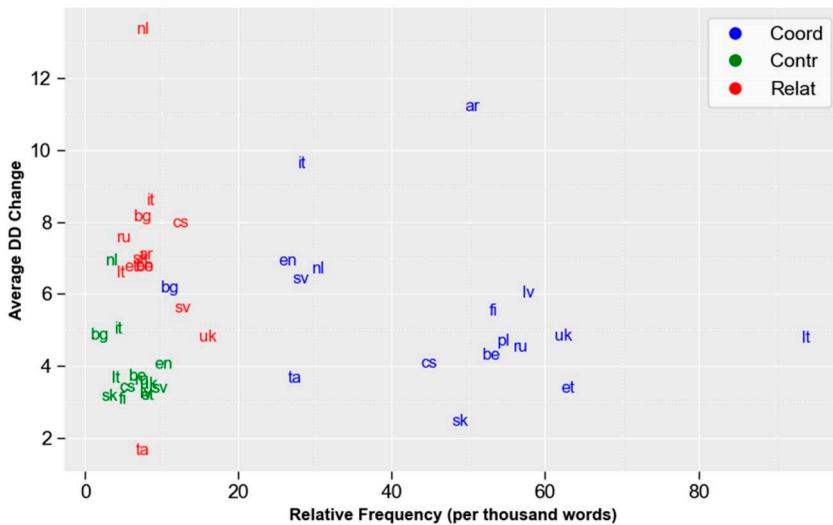
$P = 6.1e - 05 < 0.001$ ; coord vs. relat:  $W = 105$ ,  $P = 6.1e - 05 < 0.001$ ). As for the other two, while statistical test ( $W = 66$ ,  $P = 0.017$ ) shows that the relative frequencies of relative clauses are higher than those of control constructions, the differences displayed in the figure are not very significant. This might be due to the scarcity of data in the overlapping of two groups. In all, coordination is the major type of construction that determines the difference between the overall MDDs measured within two annotation schemes at corpus level. The high frequency of coordination aligns with its ubiquitous feature in language reported in the

literature (Lohmann 2014: 15). According to Mel'čuk (1988), coordination constitutes half of all types of binary linguistic relations between words, which should be regarded as a large family per se, rather than just a single type of relation. However, detailed inspection of the frequencies also suggests that more frequent constructions might weaken the impact of relatively less common phenomena.

The vertical axis of Fig. 5, on the other hand, provides information about the average DD changes of three constructions. They display a different order compared with relative frequency. The relative clause group is

**Table 3.** The overall MDDs under eight controlled situations.

Language	[−coord −contr −relat]	[+coord −contr −relat]	[−coord +contr −relat]	[−coord −contr +relat]	[+coord +contr −relat]	[−coord −contr +relat]	[+coord −contr −relat]	[−coord +contr +relat]	[+ coord + contr + relat]
Arabic	2.869	3.323	2.869	2.885	3.323	3.335	2.885	3.335	3.335
Belarusian	2.169	2.470	2.181	2.185	2.479	2.483	2.197	2.492	2.333
Bulgarian	2.266	2.312	2.271	2.281	2.318	2.327	2.287	2.327	2.333
Czech	2.390	2.670	2.396	2.404	2.674	2.681	2.410	2.685	2.685
Dutch	2.928	3.047	2.943	2.929	3.061	3.047	2.944	3.061	3.061
English	2.532	2.651	2.548	2.543	2.665	2.662	2.559	2.676	2.676
Estonian	2.545	2.972	2.551	2.542	2.974	2.968	2.548	2.970	2.970
Finnish	2.289	2.469	2.294	2.289	2.472	2.469	2.294	2.472	2.472
Italian	2.503	2.710	2.514	2.521	2.721	2.726	2.532	2.736	2.736
Latvian	2.417	2.629	2.424	2.417	2.634	2.629	2.424	2.634	2.634
Lithuanian	2.585	3.028	2.589	2.585	3.030	3.027	2.590	3.029	3.029
Polish	2.214	2.493	2.214	2.214	2.493	2.493	2.214	2.493	2.493
Russian	2.382	2.736	2.392	2.390	2.742	2.741	2.399	2.747	2.747
Slovak	2.062	2.295	2.066	2.069	2.298	2.301	2.073	2.304	2.304
Swedish	2.473	2.588	2.482	2.484	2.595	2.597	2.493	2.605	2.605
Tamil	2.398	2.439	2.398	2.392	2.439	2.433	2.392	2.433	2.433
Ukrainian	2.425	2.771	2.435	2.439	2.777	2.779	2.448	2.784	2.784



**Figure 5.** The relative frequencies and DD changes of the three constructions.

surprisingly found to be the highest in this dimension, followed by coordination, and lastly by the control construction. The difference between the first two groups is approaching the significance level (relat vs. coord:  $W = 76, P = 0.07$ ), and those between the others are significant (coord vs. contr: DD change:  $W = 98, P = 0.001$ ; relat vs. contr: DD change:  $W = 78, P < 0.001$ ).

However, although the amount of DD change induced by relative clauses per sentence is the largest, compared with the huge number of dependencies in a treebank, their contribution to the overall value at corpus level seems neglectable. Taken together, it can be concluded that the factor of relative frequency plays a more decisive role in determining the MDD at corpus level at least in this case.

Thus far, while we have explained the major difference between the overall MDDs in two annotation schemes at corpus level, we are still interested in other influencing factors determining the MDDs at sentence level, even if they do not contribute significantly to the overall MDD at corpus level.

### 4.3 Factors affecting values of MDDs at sentence level

#### 4.3.1 Stochastically intervening words

One potential factor that plays a part in determining MDDs at sentence level is the stochastically intervening words. The examples in Fig. 2 just showed the simplest cases with a few words. Nevertheless, once there are other constituents intervening between the two ends of the dependencies, then the new dependencies might raise the overall MDD. For instance, in a simple sentence, *John and Mary got married*, the DD of the added dependency between *Mary* and *married* is 2,

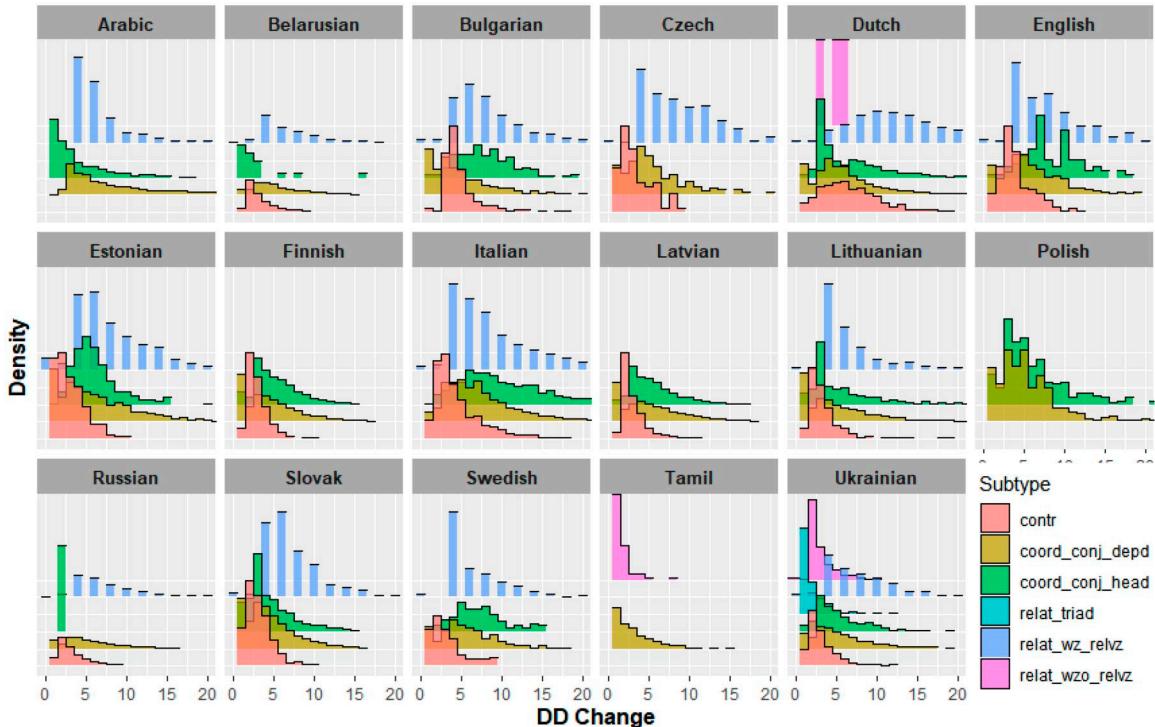
while if we expand it to *John and Mary, who had known each other for a long time, got married yesterday*, the DD change becomes 10, which is much higher than the former. Note that the size of the intervening constituents cannot be predicted from structural analysis but determined by the content that the addresser expresses, thereby being purely random.

How does it determine the direction of MDD change then? For mathematical rationales, whether MDD increases or decreases depends on the change of numerator (total dependency distance) and denominator (the number of dependencies) simultaneously according to formula (3). Therefore, whether the overall MDD is supposed to increase or decrease depends on whether the change in DD is higher or lower than the original total MDD, which can be easily proven. Since the original MDDs are generally between 2 and 3 as illustrated in Fig. 4, we can derive the following rule. In doing so, the determination of MDD change is reduced to the assessment of DD change.

#### 4.3.1.1 A simple criterion for determining whether MDD increases or decreases

In general, as long as the dependency distance of the added dependency is equal to or smaller than the threshold value 2, then it would lead to the decrease of the overall MDD; conversely, if it exceeds 2, it will cause an increase in the MDD.

The DD changes caused by enhanced dependencies across all texts were calculated and grouped according to the subtypes of constructions, following the method introduced in Section 3.2. A frequency count was then performed to obtain the probability distributions. Consequently, Fig. 6 demonstrates the overall



**Figure 6.** The distributions of DD changes caused by all subtypes of constructions.

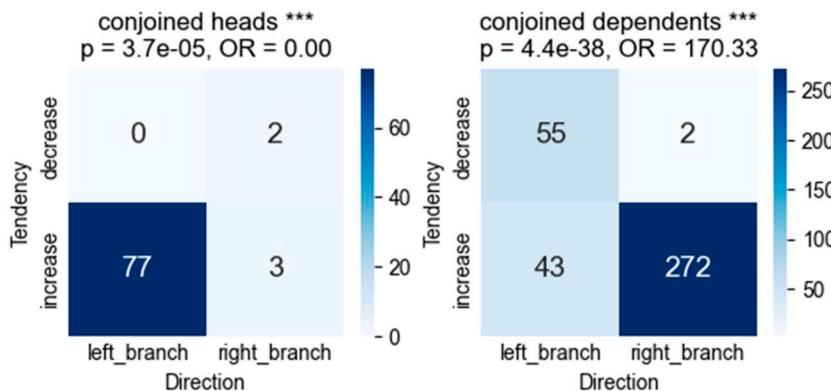
distributions across different lengths of DD changes for each construction. As shown above, most values of DD changes are larger than 2, which explain the general tendency of MDD change in EUD to increase when compared with BUD. Indeed, it has to be admitted that being below the threshold of 2 is somewhat too hard a condition. However, from Fig. 6, we are still able to identify a certain proportion of cases that leads to the decrease of MDDs, namely, those with a DD change of 1. More precisely, it is possible for all three broad types of constructions to have decreased MDD theoretically though the proportion is small, which supports Yih, Yan, and Liu's (2022) preliminary finding in smaller treebanks. It provides evidence for our argument that the impact of annotation scheme on the values of textual measures such as MDD is not deterministic but probabilistic. Moreover, our results also reveal that due to the averaging nature of MDD, it is hard for the reanalysis of single, especially rare, phenomenon to cause huge changes to the overall value. In other words, if the difference between two annotation schemes is not substantial enough, the value of MDD would not be influenced significantly. This could be a distinguishing feature of MDD compared to other similar but intrinsically different, non-averaged measures, such as 'dependency length' or 'total dependency distance' (Futrell, Mahowald, and Gibson 2015).

#### 4.3.2 Word order (with focus on coordination)

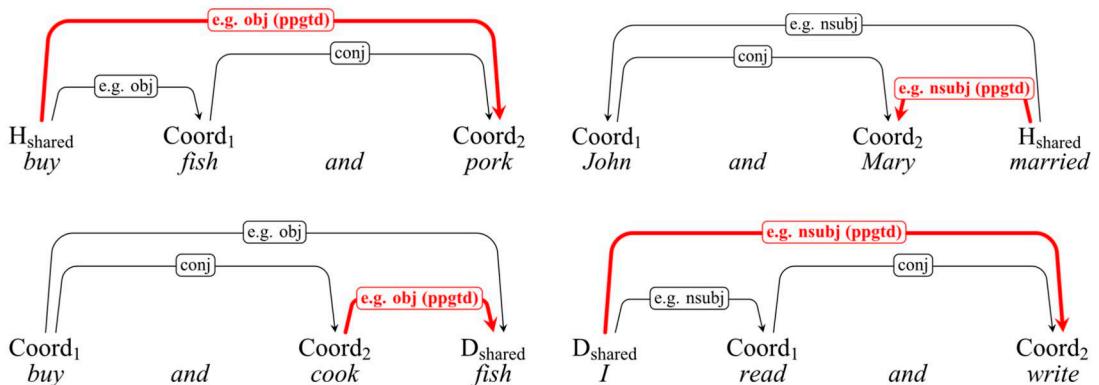
In this section, we turn to explore another factor behind MDD changes, namely, word order, by focusing exclusively on coordination, since it was found to play the most important role due to its high frequency in all treebanks.

Figure 7 shows that in English, whether MDD increases or decreases is correlated with the direction of dependent in coordinate structures. In the case of coordination with conjoined dependents, left branching has a higher odds ratio to decrease compared with right branching. Conversely, for coordination with conjoined heads, right branching has a higher odds ratio. The rationale behind these is that a shorter propagated dependency is added to make a triad for certain branching orders. If the opposite order is taken, the added dependency would be the long one, as shown in Fig. 8. Note that even in those configurations where the added dependency is the longer edge in the triad, there is still possibility for it to decrease the total MDD, as long as its DD value is two, smaller than the average level. This happens when there is no coordinator between coordinands or there is simply a comma, which would be omitted during the calculating MDD.

For the cases in all seventeen languages, see Appendix 2. For coordination with conjoined dependents, all except Tamil, the correlation holds



**Figure 7.** Correlation heatmaps for the tendency of change of MDD and direction in English (\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , P-values calculated by chi-squared tests).



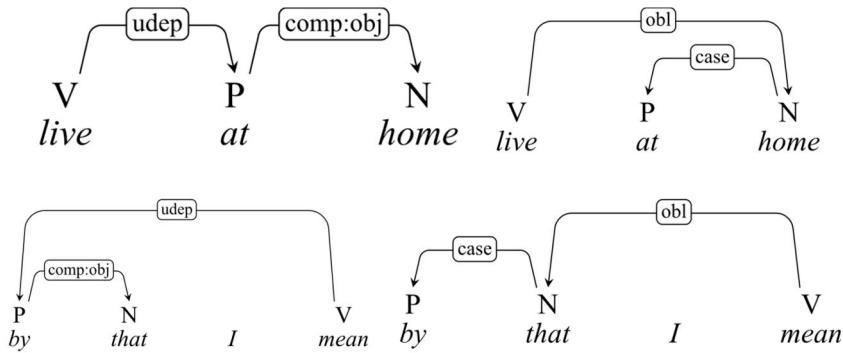
**Figure 8.** Two types of coordination with dependents on different sides (upper left: right branching with conjoined dependents; upper right: left branching with conjoined dependents; lower left: right branching with conjoined heads; lower right: left branching with conjoined heads).

significantly (\*\*\*) with a  $P$ -value smaller than 0.001. As for coordination with conjoined heads, leaving aside Czech, Russian, and Tamil with few data in such cases, most languages also have high  $p$ -values tested by chi-squared tests, and high odds ratios.

Our analysis can also be used to explain Yan and Liu's (2022) previous finding that the MDD calculated within the framework of SUD is higher than that of UD. The differences between SUD and UD can be decomposed into the distinct treatment of four constructions, where UD takes a content-head approach whereas SUD a function-head approach. All four distinctive constructions also fall into the triad local configuration proposed in this article. Here we take the prototypical adposition construction as an example,

which can be seen as a triad (V/P/N<sup>14</sup>) composed of three elements, a verb (V), an adposition (P), and a noun (N). Figure 9 shows all four possible linearizations.<sup>15</sup> Note that the MDDs of mirror orders are equal (e.g. MDD of V-P-N = MDD of N-P-V).

If the construction in a language has a linear word order of V-P-N (or its mirror order N-P-V), then apparently MDD within SUD is lower than that within UD. However, once the word order becomes V-N-P (or its mirror order P-N-V), that is, the adposition does not occur between the verb and the noun, then the UD framework has a shorter MDD with regard to SUD. Hence, what the results actually reveal is the word order universality that natural languages have the tendency to place the adposition between verbs



**Figure 9.** The treatment of V/P/N patterns in SUD and UD (upper left: VPN configuration in SUD; upper right: VPN configuration in UD; lower left: PNV, the mirror configuration of VNP, in SUD; lower right: PNV, the mirror configuration of VNP, in UD).

and nouns (Dik 1997). In other words, V-P-N is universally more frequent than the V-N-P pattern, which forms the basis for the MDD of SUD to be larger than that of UD. We show that Yan and Liu's finding at corpus level is also a statistical result with mixed possibilities at sentence level.

Similarly, for the case of coordination in the present study, what we find is in fact a word order universal that the left-branching in coordination with conjoined heads and right branching in coordination with conjoined dependents are more frequent than the corresponding opposite cases. It is this correlation that lead to the increase of  $MDD_{EUD}$  compared with  $MDD_{BUD}$ , at least in terms of coordinate structures.

Thus far, we have analyzed the factors behind the difference of MDD values in two annotation schemes and reached a number of findings. The remaining question is then: What kind of view should we hold regarding choosing between annotation schemes? It will be addressed in the next section.

#### 4.4 Revisiting the relation between annotation scheme and textual measures

This section intends to address the core question of this paper: what view should we hold regarding the relationship between annotation scheme and MDD?

Yih, Yan, and Liu (2022) previously drew a parallel between the relationship of  $MDD_{EUD}$  and  $MDD_{BUD}$  and that of different temperature scales (e.g. Fahrenheit and Celsius). However, the findings of this paper compel us to revise this view. It has been indicated that the final change in MDD induced by the annotation scheme is probabilistic rather than deterministic, collaboratively decided by the interaction of multiple factors. Since an annotation scheme involves numerous syntactic

constructions, the change in MDD is almost unpredictable before measurement. Therefore, there is no functional relation or conversion formula between  $MDD_{EUD}$  and  $MDD_{BUD}$  (or between  $MDD_{UD}$  and  $MDD_{SUD}$ ) in analogy with different scales of temperature.

In this article, we propose the novel view that various versions of MDDs calculated based on different annotation schemes ( $MDD_{EUD}$ ,  $MDD_{BUD}$ , among others) should better be viewed as distinct textual measures. Given that the temperature scale metaphor does not hold, we draw new analogies to the differences between MDD and total dependency distance, or between the absolute frequency and normalized frequency of a particular structure. Despite the terminologies appearing similar at first glance, these measures are fundamentally different from one another in terms of the algorithm, the formulae or the intermediate preprocessing procedure. In our case of MDD variants, the discrepancies are primarily caused by the annotation scheme which specifies the output representation of syntactic parsing, although the final numerical differences are relatively small. One piece of evidence supporting our view comes from the line chart of Fig. 4: sorting the MDD values of different languages before and after enhancement results in different orders. Likewise, Yan and Liu's (2022) data show similar pattern in their comparison of the UD and SUD annotation schemes. The underlying rationale is that if the two measures are identical operationalizations of a construct, their rankings should be consistent. However, the discrepancy in the order suggests that the two measures are in fact capturing different constructs.

Based on the perspectives above, the following two implications can be derived. First, if the measure of MDD is initially designed to reflect syntactic complexity (Liu 2008), understood as difficulty or

sophistication, or further used as a measure of learners' proficiency or degree of language attrition, then an important objective should be to choose between different versions of MDDs. This is an empirical issue of the validation of measures (e.g. see Drown *et al.* 2024a, b for a full validation process of two measures of vocabulary knowledge). One might conduct criteria validity by taking an external variable as golden standard. Therefore, a potential area for future research is to follow studies like Niu and Liu (2022), who compared three complexity measures of sentences (MDD, total dependency length, and sentence length) with the reaction time in sentence comprehension.

A further issue is whether the measure of MDD could guide the selection of the annotation scheme. Recall that in the literature some held that SD/SUD is preferred over UD due to their lower statistical MDD based on the principle of Dependency Distance Minimization (DDM) (Gerdes *et al.* 2018; Osborne and Gerdes 2019). However, there seem to be several concerns with this argument. First, the DDM principle is a linguistic universal reflecting language complexity, originally discovered based on the comparison between natural languages and random languages (Liu 2008), with the MDD under a specific annotation framework chosen as a measure of working memory. This principle inherently holds, irrespective of the choice of annotation scheme (Yan and Liu 2022: 425). Therefore, while choosing syntax-oriented annotation schemes superficially lowers the MDD value, it does not reduce syntactic complexity or working memory itself; rather, it simply substitutes one measure for another. Second, if the MDDs are deliberately designed to be minimized to obtain a framework with the lowest MDD, then it causes logical circularity. Moreover, even if an annotation scheme aimed at minimizing the MDD were to be pursued, theoretically, one cannot find a consistent scheme for any language under any circumstances. Our results in Section 4.2 show that whether the syntactic analysis of certain phenomenon contributes to the increase or decrease of the overall MDD is determined interactively by multiple factors, revealing that there is a trade-off between the consistency of annotation and the change in MDD.

However, we believe that Gerdes and his colleagues' argument can still serve as a legitimate reason for selecting SD/SUD as a preferable framework for syntactic-oriented analysis. Yet it should not be interpreted from the perspective of the DDM principle, but approached from a different angle. What Gerdes *et al.* in fact reveal is that  $MDD_{SD/SUD}$ , compared with

$MDD_{UD}$ , is a more valid measure of syntactic complexity, although their conclusion is drawn based on traditional intuitive judgments of certain syntactic constructions rather than psycholinguistic experiments. This matches the research paradigm mentioned in the previous paragraph, that is, pursuing a better operationalized measure of a construct, which justifies the choice of an annotation scheme. Nevertheless, this does not rule out the possibility of using other frameworks with higher MDD values, such as UD and EUD, for other appropriate purposes. As Osborne and Gerdes (2019) themselves pointed out, one should 'allow the individual researcher to choose which of the two sets of treebanks best matches his or her goals' (p. 24), which is also in line with Yih's (2023) idea the choice between annotation schemes is not a matter of correctness (right or wrong), but rather of appropriateness (good or bad).

## 5. Conclusion

Following previous studies, the present article further explored the effect of annotation scheme on the textual measure of MDD within a dependency grammar framework. We compared the treebanks of 17 languages which have both tree (BUD) and graphic (EUD) representations in the UD project. Empirical results show that the overall MDDs in the EUD representation are statistically higher than those in the basic UD representation at corpus level, while at sentence level, for all the three major distinctive constructions, there are cases where they increase or decrease MDD. Specifically, the EUD analysis of coordinate structures contributes most to the increase of MDDs due to its higher relative frequencies in the treebanks. We also explored the microscope factors that lead to the changes, including stochastically intervening words and word order. Finally, the view on the relation between annotation schemes and textual measures was discussed. We proposed that the MDDs under different annotation schemes should be regarded as different textual measures, rather than different scales of the same measure. As for which measure is more appropriate to use in practice, it depends on the results of external validation. Moreover, Dependency Distance Minimization should not be used as evidence for selecting the annotation scheme.

This study undoubtedly also has some limitations. For instance, the potential impact of unbalanced treebank sizes has not been fully explored. This is partially due to the inherent constraints of UD data, which are the only large-scale cross-linguistic corpora accessible

to us. Another limitation comes from the errors in the annotation of treebanks. As seen in the previous sections, certain specific structures related to enhanced dependencies in some languages have not been accurately annotated, which weakens the accuracy of our results to certain extent. The addition of more high-quality data in the future would effectively help address the issues mentioned above.

For future studies, one may create and compare more variants of MDD by decomposing annotation schemes into more micro-constructions and manually calibrating the analysis of each construction. In addition, the effect of annotation schemes on other syntactic complexity measures beyond MDD(s) is also worth exploring, such as dependency direction, network metrics of syntactic networks, as well as other widely used measures in second language research (e.g. mean length of utterance). Furthermore, it is worth investigating whether there are measures that remain reliable across different annotation schemes.

## Author contributions

Tsy Yih (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Visualization, Writing—original draft, Writing—review & editing) Haitao Liu (Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review & editing)

## Acknowledgements

The first author would like to thank Chengdu AG (All Gamers), especially Yinuo, for their emotional support during the writing of this paper, and his three cats—Miaomiao, Mimi, and Monki—for appearing in the example sentence.

## Funding

This work was supported by the National Social Science Fund of China (20CYY030), the Postdoctoral Fellowship

Program of CPSF (GZB20240559), and the China Postdoctoral Science Foundation (2024M762392).

## Notes

1. For instance, it can be used as a valid indicator of second language learners' proficiency (Ouyang and Jiang 2018; Li and Yan 2021) or the degree of older adults' language attrition (Liu, Zhao, and Bai 2021).
2. The term UD could be polysemous. In one sense, it refers to a specific annotation scheme in contrast with SD or others. In another sense, it denotes the tabular data exchange format, CoNLL-U, which is sometimes also referred to as the UD format in abbreviation. The last sense stands for an initiative of building dependency treebanks following the UD annotation scheme and presented in the CoNLL-U format (Zeman *et al.* 2017), which already has 283 treebanks of 161 languages till v2.14 (<https://universaldependencies.org/>).
3. For more information, the reader is referred to <https://universaldependencies.org/u/overview/enhanced-syntax.html>.
4. The term 'coordinand' is generally used in the typological literature (Haspelmath 2004). It is also referred to as 'conjunction' in generative or theoretical linguistics. In this study, we consistently use the term 'coordinand'.
5. All the treebanks are available from <https://universaldependencies.org/>.
6. Although some treebanks claim to have enhanced dependencies in their descriptions, we found that they are simply the copy of basic dependencies without additional links.
7. <https://scipy.org/>
8. <https://ggplot2.tidyverse.org/>
9. <https://www.r-project.org/>
10. <https://matplotlib.org/>
11. <https://seaborn.pydata.org/>
12. Punctuations are generally not included in the calculation of MDD.
13. There are in fact two approaches to calculating the MDD of the whole treebank. An alternative approach is Lundholm Fors, Fraser, and Kokkinakis's (2018) method, where the MDDs are first calculated at sentence level and then averaged to obtain the overall MDD of the treebank. In this paper we adopted Liu's approach which puts all dependencies within a treebank together. The Liu's method is preferred because it avoids the influence of sentence length (Jiang and Liu 2015), which might cause a problem for Lundholm Fors *et al.*'s approach.
14. We use slash to denote a triad before linearization (e.g. V/P/N = V/N/P, etc.), and hyphen to denote a pattern with certain order (e.g. V-P-N ≠ V-N-P).
15. The rest two logical orders P-V-N and N-V-P are hardly attested because generally the function of the adposition is to help identify the nominal argument, and together they always form a phrase first.

## Appendix 1 Algorithms for identifying all types of constructions in CoNLL-U format treebanks (pseudo-codes)

### Algorithm 1 Identifying coordination, control constructions, and relative clauses with full triad

```

for sent in sents: # 'sent' stands for sentence
    for tokenc in sent if LEN(DEPS_OF(tokenc)) > 1: # Start searching with token c, the node with multiple heads in the triad
        for dep3 in NON_ROOT_DEPS_OF(tokenc): # 'dep' stands for dependency (relation)
            tokena ← FIND_HEAD_OF_ALONG(tokenc, dep3) # Find the token a, which governs two tokens in the triads
            for tokenb in sent:
                if IS_HEAD_OF(tokena, tokenb) and IS_HEAD_OF(tokenb, tokenc): # Find the intermediate node, token b, in the triad
                    for dep1 in NON_ROOT_DEPS_OF(tokenb):
                        if HEAD_OF(dep1) = HEAD_OF(dep3):
                            for dep2 in Non_ROOT_DEPS_OF(tokenb):
                                if HEAD_OF(dep2) = Id_OF(tokenb):
                                    # ↑ Thus far, having found a triad
                                    if TYPE_OF(dep2) = conj and TYPE_OF(dep1) ≠ conj:
                                        return 'coord_conj_depd'
                                    if TYPE_OF(dep1) = conj and TYPE_OF(dep2) ≠ conj:
                                        return 'coord_conj_head'
                                    if TYPE_OF(dep1) = xcomp and TYPE_OF(dep2) ≠ conj:
                                        return 'contr'
                                for tokend in sent: # Find the subordinate head, token d, in the case of relative clause
                                    for dep4 in Non_ROOT_DEPS_OF(tokend):
                                        if HEAD_OF(dep4) = Id_OF(tokenb) and TYPE_OF(dep4) = acl: relcl:
                                            return 'relat_triad'
```

### Algorithm 2. Identifying relative clauses with relativizers (but not triad)

```

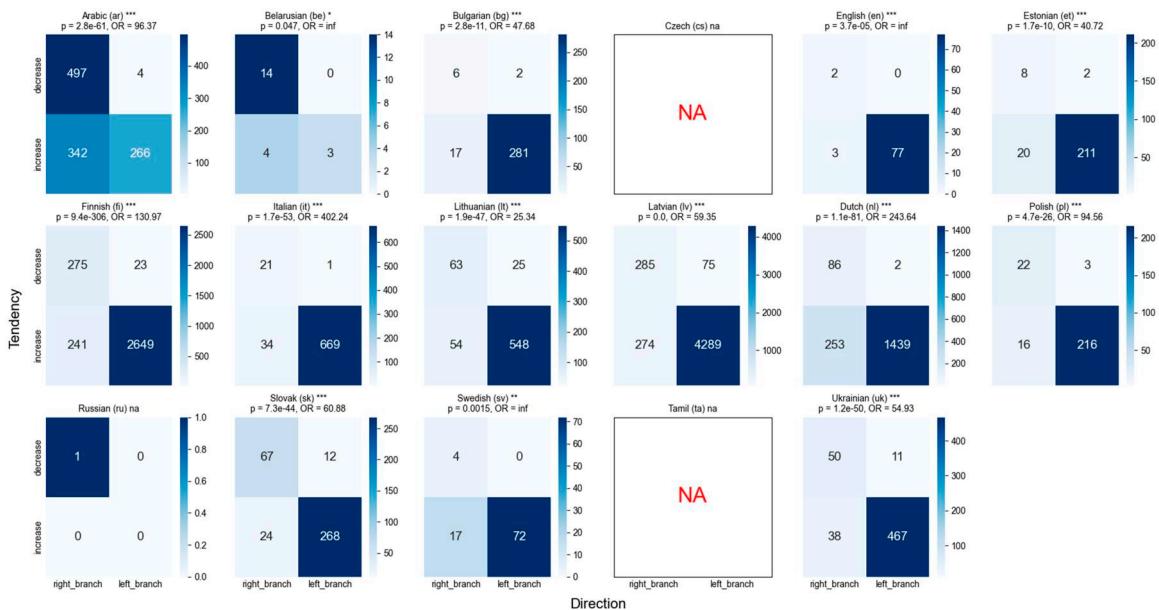
for sent in sents:
    for tokenb in sent if LEN(DEPS_OF(tokenb)) > 1: # Start searching with token b, the antecedent
        for tokend in sent:
            for dep4 in Non_ROOT_DEPS_OF(tokend):
                if HEAD_OF(dep4) = Id_OF(tokenb) and TYPE_OF(dep4) = acl: relcl:
                    for tokena in sent:
                        if IS_HEAD_OF(tokend, tokena) and IS_HEAD_OF(tokenb, tokena):
                            for dep1 in Non_ROOT_DEPS_OF(tokenb):
                                if HEAD_OF(dep1) = Id_OF(tokena):
                                    for tokenc in sent:
                                        for dep2 in Non_ROOT_DEPS_OF(tokenc):
                                            if HEAD_OF(dep2) = Id_OF(tokenb) and TYPE_OF(dep2) = ref:
                                                return 'relat_wz_relvz'
```

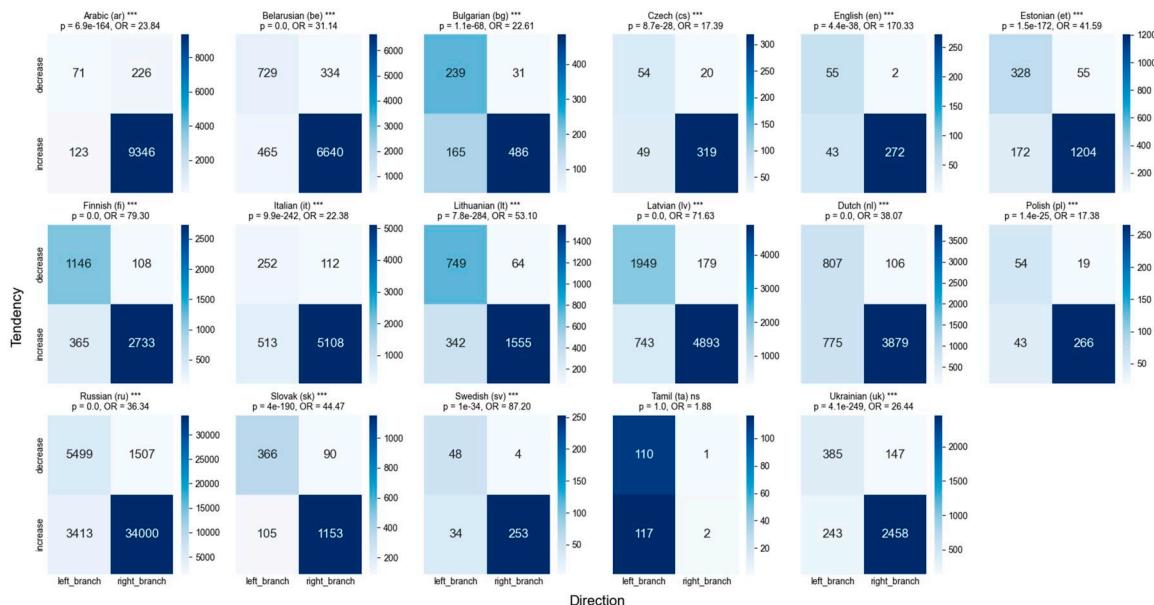
**Algorithm 3.** Identifying relative clauses without relativizers

```

for sent in sents: # 'sent' stands for sentence
    for tokenb in sent if LEN(DEPS_OF(tokenb)) > 1: # Start searching with token b, the antecedent
        for dep1 in Non_ROOT_DEPS_OF(tokenb):
            tokend  $\leftarrow$  FIND_HEAD_OF_ALONG(token, dep3)
            for dep4 in Non_ROOT_DEPS_OF(tokend):
                if HEAD_OF(dep4) = Id_OF(tokenb) and TYPE_OF(dep4) = acl: relcl:
                    if not any(TYPE_OF(dep2) = ref and HEAD_OF(dep2) = Id_OF(tokenb) for token3 in sent for dep2 in DEPS_OF(tokend)):
                        return 'relat_wzo_relvz'

```

**Appendix 2 Correlation heatmaps for the cases of coordination with conjoined dependents/heads in all 17 languages****Figure A1.** Correlation heatmaps for the cases of coordination with conjoined heads in all seventeen languages.



**Figure A2.** Correlation heatmaps for the cases of coordination with conjoined dependents in all seventeen languages.

## References

- de Marneffe, M.-C., and Manning, C. D. (2008) ‘The Stanford Typed Dependencies Representation’, in J. Bos *et al.* (eds) *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Manchester: Coling 2008 Organizing Committee.
- de Marneffe, M.-C. *et al.* (2014) ‘Universal Stanford Dependencies: A Cross-linguistic Typology’, in N. Calzolari *et al.* (eds) *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 26–31. Reykjavik: European Language Resources Association (ELRA).
- Dik, S. C. (1997) *The Theory of Functional Grammar*, 2nd ed. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110218367>
- Drown, L. *et al.* (2024a) ‘Validation of Two Measures for Assessing English Vocabulary Knowledge on Web-based Testing Platforms: Brief Assessments’, *Linguistics Vanguard* 9: 99–111.
- Drown, L. *et al.* (2024b) ‘Validation of Two Measures for Assessing English Vocabulary Knowledge on Web-based Testing Platforms: Long-form Assessments’, *Linguistics Vanguard* 9: 113–24.
- Ferrer-i-Cancho, R., and Gómez-Rodríguez, C. (2021) ‘Anti dependency Distance Minimization in Short Sequences. A Graph Theoretic Approach’, *Journal of Quantitative Linguistics*, 28: 50–76. <https://doi.org/10.1080/09296174.2019.1645547>
- Futrell, R., Mahowald, K., and Gibson, E. (2015) ‘Large-scale Evidence of Dependency Length Minimization in 37 Languages’, *PNAS*, 112: 10336–41. <https://doi.org/10.1073/pnas.1502134112>
- Futrell, R., Levy, R. P., and Gibson, E. (2020) ‘Dependency Locality as an Explanatory Principle for Word Order’, *Language*, 96: 371–412. <https://doi.org/10.1353/lan.2020.0024>
- Gerdes, K. *et al.* (2018) ‘SUD or Surface-Syntactic Universal Dependencies: An Annotation Scheme Near-isomorphic to UD’, in Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster (eds) *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 66–74. Brussels: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6008>
- Hao, Y., Wang, X., and Lin, Y. (2022) ‘Dependency Distance and its Probability Distribution: Are They the Universals for Measuring Second Language Learners’ Language Proficiency?’, *Journal of Quantitative Linguistics*, 29: 485–509. <https://doi.org/10.1080/09296174.2021.1991684>
- Haspelmath, M. (2004) ‘Coordinate Structures: An Overview’, in M. Haspelmath (ed.) *Coordinate structures*, pp. 3–39. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.58>
- Hovy, D. (2020) *Text Analysis in Python for Social Scientists: Discovery and Exploration*. Cambridge: Cambridge University Press.
- Hovy, D. (2022) *Text Analysis in Python for Social Scientists: Prediction and Classification*. Cambridge: Cambridge University Press.
- Hudson, R. (1995) ‘Measuring Syntactic Difficulty’, unpublished manuscript. <https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>, accessed 8 May 2024.
- Hudson, R. (2007) *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
- Jiang, J., and 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. <https://doi.org/10.1016/j.langsci.2015.04.002>
- Jiang, J., and Ouyang, J. (2018) ‘Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition’, In J. Jiang and H. Liu (eds) *Quantitative Analysis of Dependency Structures*, pp.

- 167–90. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110573565-009>
- Jing, Y., and Liu, H. (2015) ‘Mean Hierarchical Distance Augmenting Mean Dependency Distance’, in *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 161–70. Uppsala: Uppsala University. <https://aclanthology.org/W15-2119/>
- Johnson, W. (1944) ‘Studies in Language Behavior: A Program of Research’, *Psychological Monographs*, 56: 1–15.
- Li, W., and Yan, J. (2021) ‘Probability Distribution of Dependency Distance Based on a Treebank of Japanese EFL Learners’ Interlanguage’, *Journal of Quantitative Linguistics*, 28: 172–86.
- 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: 159–91.
- Liu, H. (2010) ‘Dependency Direction as a Means of Word-order Typology: A Method Based on Dependency Treebanks’, *Lingua*, 120: 1567–78. <https://doi.org/10.1016/j.lingua.2009.10.001>
- Liu, H. (2022) *Dependency Relations & Language Networks* (In Chinese). Beijing: Science Press.
- Liu, H., Hudson, R., and Feng, Z. (2009) ‘Using a Chinese Treebank to Measure Dependency Distance’, *Corpus Linguistics and Linguistic Theory*, 5: 161–74. <https://doi.org/10.1515/CLLT.2009.007>
- Liu, H., Xu, C., and Liang, J. (2017) ‘Dependency Distance: A New Perspective on Syntactic Patterns in Natural Languages’, *Physics of Life Reviews*, 21: 171–93. <https://doi.org/10.1016/j.plrev.2017.03.002>
- Liu, J., Zhao, J., and Bai, X. (2021) ‘Syntactic Impairments of Chinese Alzheimer’s Disease Patients from a Language Dependency Network Perspective’, *Journal of Quantitative Linguistics*, 28: 253–81. <https://doi.org/10.1080/09296174.2019.1703485>
- Lohmann, A. (2014) *English Coordinate Constructions: A Processing Perspective on Constituent Order*. Cambridge: Cambridge University Press.
- Lundholm Fors, K., Fraser, K., and Kokkinakis, D. (2018) ‘Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment’, in A. Ugon *et al.* (eds) *Building Continents of Knowledge in Oceans of Data: The Future of Co-created eHealth*, pp. 705–709. Amsterdam: IOS Press.
- Mel’čuk, I. A. (1988) *Dependency Syntax: Theory and Practice*. New York: SUNY Press.
- Niu, R., and 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: 1043–62. <https://doi.org/10.1007/s10936-022-09878-4>
- Nivre, J. *et al.* (2016) ‘Universal Dependencies v1: A Multilingual Treebank Collection’, in N. Calzolari *et al.* (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 1659–66. Portorož: European Language Resources Association (ELRA).
- Osborne, T., and Gerdes, K. (2019) ‘The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD)’, *Glossa: A Journal of General Linguistics*, 4: 17. <https://doi.org/10.5334/gjgl.537>
- Ouyang, J. and Jiang, J. (2018) ‘Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners?’, *Journal of Quantitative Linguistics*, 25: 295–313. <https://doi.org/10.1080/09296174.2017.1373991>
- Schneider, G. (2024) *Text Analytics for Corpus Linguistics and Digital Humanities: Simple R Scripts and Tools*. London: Bloomsbury.
- Schuster, S., and Manning, C. D. (2016) ‘Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks’, in N. Calzolari *et al.* (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 2371–8. Portorož: European Language Resources Association (ELRA).
- Wang, Y., and Liu, H. (2017) ‘The Effects of Genre on Dependency Distance and Dependency Direction’, *Language Sciences*, 59: 135–47. <https://doi.org/10.1016/j.langsci.2016.09.006>
- Yan, J., and Liu, H. (2019) ‘Which Annotation Scheme is More Expedient to Measure Syntactic Difficulty and Cognitive Demand?’, in X. Chen and R. Ferrer-i-Cancho (eds) *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pp. 16–24. Paris: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-7903>
- Yan, J., and Liu, H. (2022) ‘Semantic Roles or Syntactic Functions: The Effects of Annotation Scheme on the Results of Dependency Measures’, *Studia Linguistica*, 76: 406–28. <https://doi.org/10.1111/stul.12177>
- Yan, J., and Liu, H. (2024) ‘Investigating the Hierarchical Order of Function Words in SUD and UD Treebanks’, *Linguistic Analysis*, 43: 629–66.
- Yih, T. (2023) ‘Theoretical and Quantitative Investigations of Coordination in Dependency Grammar’, Zhejiang University dissertation, Hangzhou.
- Yih, T., and Dai, Z. (2023) ‘UPOS-DEPREL Mismatches: Detecting Annotation Errors and Improving UD Guidelines based on Linguistic Knowledge’, in Chu-Ren Huang *et al.* (eds) *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 46–58. Hong Kong: Association for Computational Linguistics. <https://aclanthology.org/2023.paclic-1.5/>
- Yih, T., Yan, J., and Liu, H. (2022) ‘How Syntactic Analysis Influences the Calculation of Mean Dependency Distance: Evidence from the Enhanced Dependency Representation’, in S. Dita, A. Trillanes, and R. I. Lucas (eds) *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pp. 83–92. Manila, Philippines: Association for Computational Linguistics.
- Zeman, D. *et al.* (2017) ‘CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies’, in J. Hajic and D. Zeman (eds) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19. Vancouver: Association for Computational Linguistics. <https://doi.org/10.1017/10.18653/v1/K17-3001>

© The Author(s) 2025. Published by Oxford University Press on behalf of EADH. All rights reserved.  
For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).  
Digital Scholarship in the Humanities, 2025, 40, 400–418  
<https://doi.org/10.1093/lhc/fqaf003>  
Full Paper