Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand?

JIANWEI YAN & HAITAO LIU

DEPARTMENT OF LINGUISTICS, ZHEJIANG UNIVERSITY

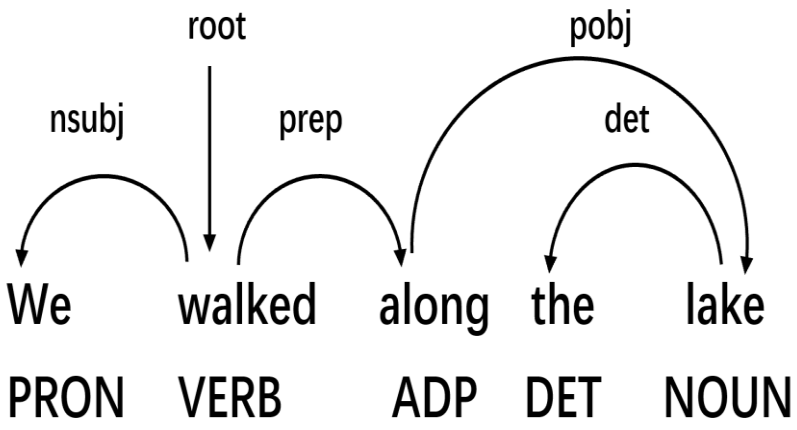
JWYAN@ZJU.EDU.CN & LHTZJU@GMAIL.COM

Outline

- Background and Motivation
- Materials and Methods
- Results and Discussion
- Conclusions and Implications

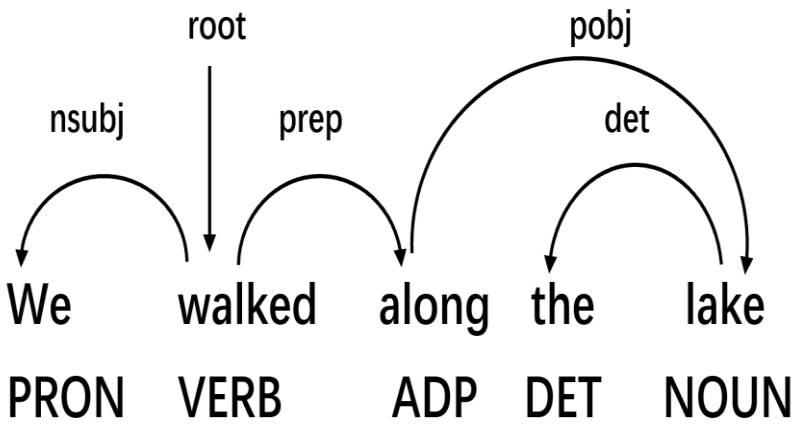
1. Background and Motivation

- The seminal work of Eléments de Syntaxe Structurale (Tesnière, 1959)
- The syntactic relations between governors and dependents within a sentence (Heringer, 1993; Hudson, 1995; Jiang and Liu, 2018).



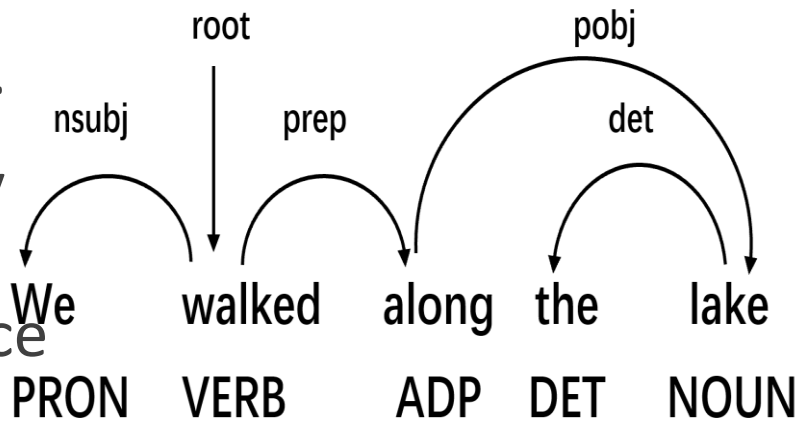
1. Background and Motivation

- Dependency distance: the linear distance of the governor and the dependent (Hudson, 1995).
- Dependency direction: the linear order of the governor and the dependent of each dependency type (Liu, 2010).



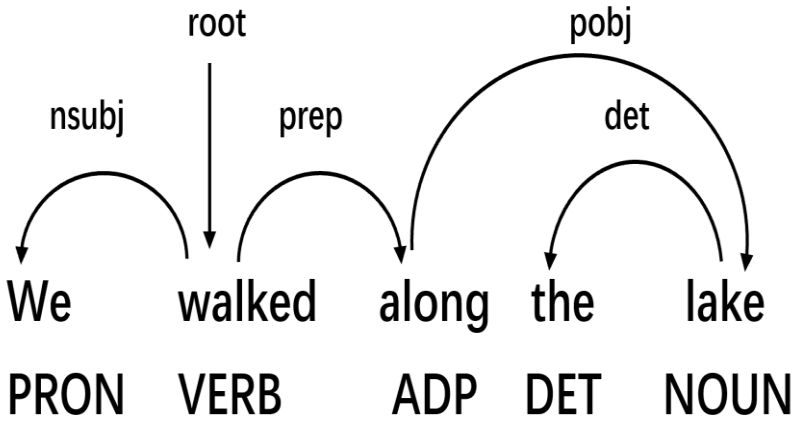
1. Background and Motivation

- Hudson (1995) proposed the definition of dependency distance.
- Based on a Romanian dependency treebank, Ferrer-i-Cancho (2004) proved that (a) the average distance of a sentence is minimized and (b) the average distance of a sentence is constrained.



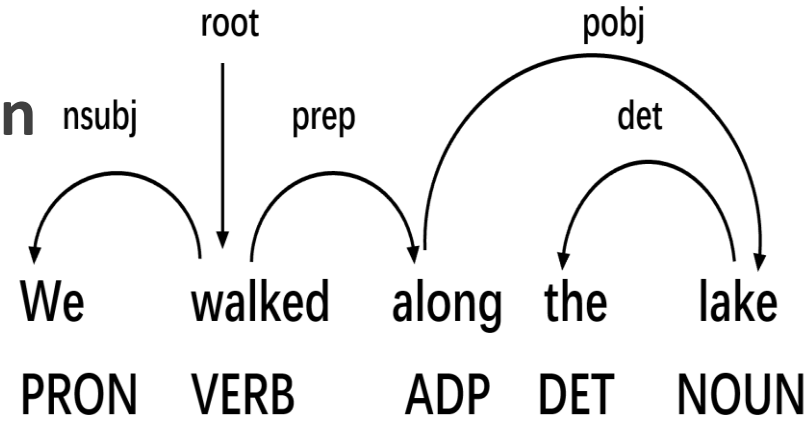
1. Background and Motivation

- Liu's (2008) empirical study on dependency distance provided a viable treebank-based approach towards the metric of syntactic complexity and cognitive constraint.
- Series of researches exploring the relationship between dependency distance and syntactic difficulty and cognitive demand have been carried out.



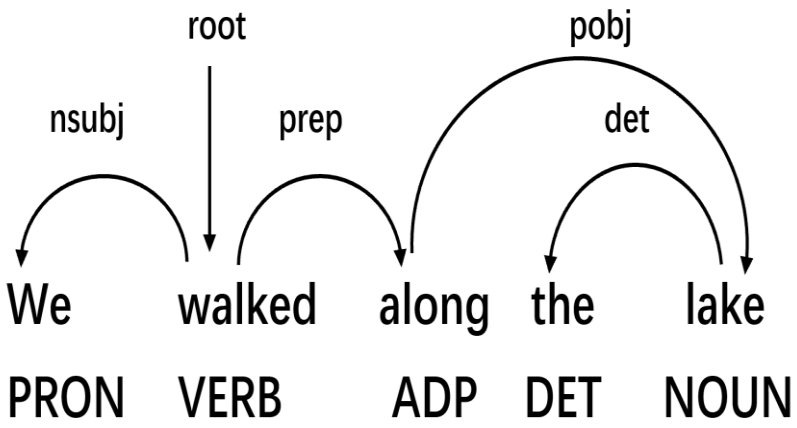
1. Background and Motivation

- The distribution of dependency distance follows the linguistic law of the Least Effort Principle (LEP) or **Dependency Distance Minimization** (DDM) (Zipf, 1965; Liu et al., 2017).
- The **mean dependency distances** (MDDs) (Liu, 2008) is an important index of memory burden, demonstrating the syntactic complexity and cognitive demand of the language concerned (Hudson, 1995; Liu et al., 2017).



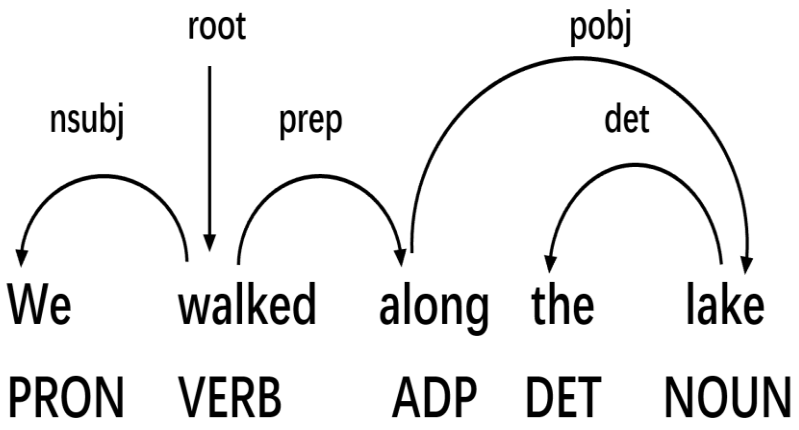
1. Background and Motivation

- There are several factors that have effects on the measurement of dependency distance, including sentence length, genre, chunking, language type, grammar, annotation scheme and so forth.
- Most of these factors have been well-investigated except the factor of annotation scheme.



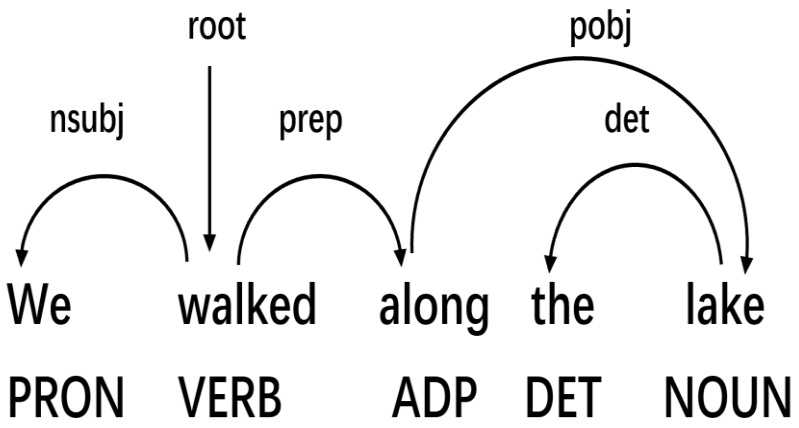
1. Background and Motivation

- Large-scale linguistic analysis under the framework of dependency grammar must be based on treebanks (annotated corpora).
- The annotated corpora must be based on specific annotation schemes, according to which the labels and associated features of linguistic units are defined (Ide and Pustejovsky, 2017).



1. Background and Motivation

- The annotation scheme of annotated resources adopted might have a great impact on the results of dependency measurements.

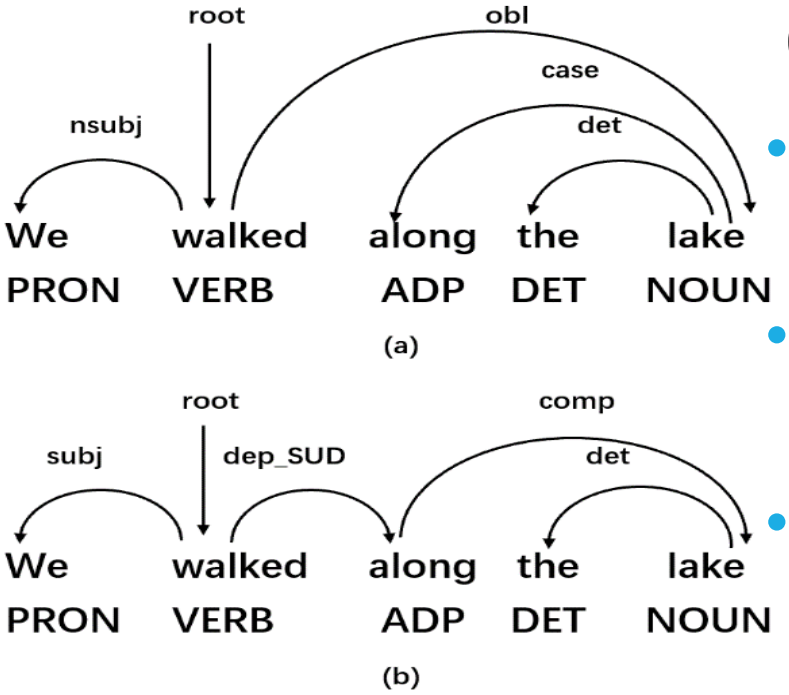


1. Background and Motivation

Research Questions:

- **Q1:** Will the probability distribution of dependency distances of natural texts change when they are based on different annotation schemes?
- **Q2:** Based on MDDs, which annotation scheme is more congruent for the measurement of syntactic complexity and cognitive demand?
- **Q3:** Which dependency types account most for the distinctions between different annotation schemes? What are the quantitative features of these dependency types?

2. Materials and Methods



- **UD:** the Universal Dependencies (Nivre, 2015)
- To hold a semantic criteria to put priorities to content words
- To maximize "crosslinguistic parallelism"
- **SUD:** the Surface-Syntactic Universal Dependencies (Gerdes et al., 2018)
- To follow the syntactic tradition
- To promote the syntactic motivations

2. Materials and Methods

- Jiang and Liu (2015) proposed several methods to compute dependency distance.
- MDD of the entire sentence can be defined as:

$$\text{MDD (}\textbf{the sentence}\text{)} = \frac{1}{n-1}\sum_{i=1}^{n-1} |\,\text{DD}_i| \qquad (1)$$

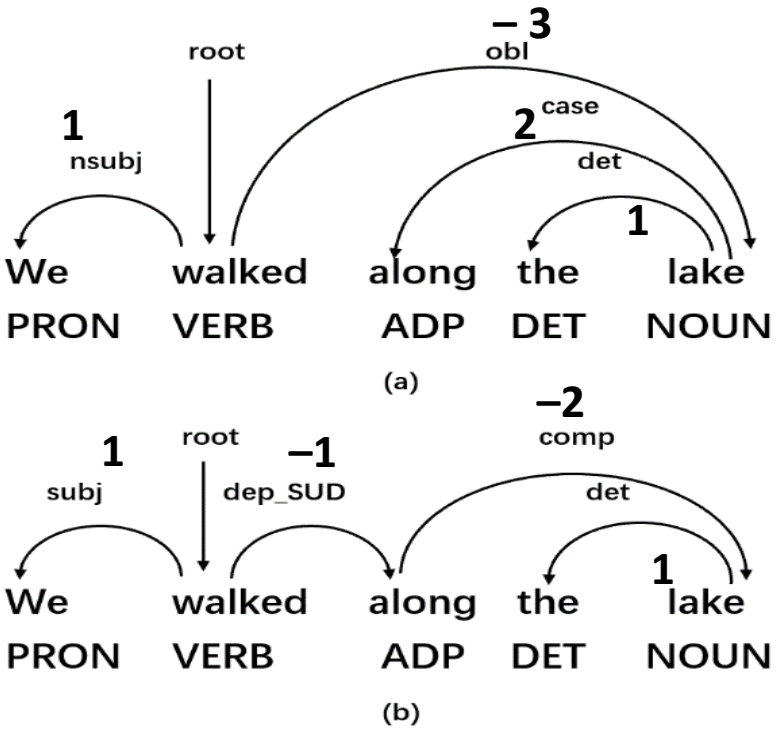
- The MDD of a treebank can be defined as:

$$\text{MDD (}\textbf{the treebank}\text{)} = \frac{1}{n-s}\sum_{i=1}^{n-s} |\,\text{DD}_i| \qquad (2)$$

- The MDD for a specific type of dependency is:

$$\text{MDD (}\textbf{dependency type}\text{)} = \frac{1}{n}\sum_{i=1}^{n} \text{DD}_i \qquad (3)$$

2. Materials and Methods



- **UD** MDD:
- (|1|+|2|+|1|+|– 3|)/4=1.75.

- **SUD** MDD:
- (|1|+|– 1|+1+|– 2|)/4=1.25.

3.1 Results and Discussion: Annotation Scheme and Probability Distribution of Dependency Distance

- The probability distribution of dependency distances of natural languages shares some regularities, including right truncated zeta (Jiang and Liu, 2015; Wang and Liu, 2017; Liu et al., 2017) and right truncated waring (Jiang and Liu, 2015; Lu and Liu, 2016; Wang and Liu, 2017).
- **Q1:** Will the probability distribution of dependency distances of natural texts change when they are based on different annotation schemes? Do they still follow the linguistic law of DDM?

3.1 Results and Discussion: Annotation Scheme and Probability Distribution of Dependency Distance

- The Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017) in UD 2.2 and SUD 2.2 projects
- Seven genres, viz. *academic writing, biographies, fiction, interviews, news stories, travel guides* and *how-to guides,* with a total amount of 95 texts.

3.1 Results and Discussion: Annotation Scheme and Probability Distribution of Dependency Distance

- Fitted dependency distances of all 95 texts of GUM to the probability distribution of right truncated zeta and right truncated waring by Altmann-Fitter.
- The determination coefficient R^2 can indicate the goodness-of-fit (Wang and Liu, 2017; Wang and Yan, 2018).

Annotation	Genre	Right truncated waring	Right truncated zeta
		R^2	R^2
UD	*Academic*	0.9844	0.8847
	Bio	0.9859	0.8980
	Fiction	0.9909	0.9119
	Interview	0.9912	0.9011
	News	0.9899	0.9149
	Voyage	0.9906	0.8999
	Whow	0.9881	0.8996
	Average	0.9887	0.9014
SUD	*Academic*	0.9964	0.9769
	Bio	0.9953	0.9808
	Fiction	0.9949	0.9770
	Interview	0.9974	0.9802
	News	0.9974	0.9833
	Voyage	0.9974	0.9787
	Whow	0.9954	0.9728
	Average	0.9963	0.9785

Table Mean values of the determination coefficient R^2 fitted to dependency distances of all genres based on UD and SUD annotation schemes.

3.1 Results and Discussion: Annotation Scheme and Probability Distribution of Dependency Distance

- Conventionally, the excellent, good, acceptable and not acceptable goodness-of-fit for determination coefficient R^2 are 0.90, 0.80, 0.75 and less than 0.75, respectively.
- The frequencies of dependency distances based on both UD and SUD treebanks can well capture the models of right truncated waring and right truncated zeta with a good coefficients of determination R^2.

3.1 Results and Discussion: Annotation Scheme and Probability Distribution of Dependency Distance

- The probability distributions of dependency distances of natural texts based on both UD and SUD annotation schemes share similar power law distribution.
- The probability distributions of dependency distances of all texts based on both UD and SUD follow the same regularity, supporting the Least Effort Principle (LEP) (Zipf, 1965) or the linguistic law of DDM (Liu, 2008; Futrell et al., 2015; Liu et al., 2017).

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- The relationship between dependency distance and syntactic difficulty and cognitive demand have been exploited by many studies, including assessing first language acquisition (Ninio, 2011, 2014), second language learning (Ouyang and Jiang, 2018; Jiang and Ouyang, 2018), syntactic development of deaf and hard-of-hearing students (Yan, 2018), etc.
- **Q2:** Based on MDDs, which annotation scheme is more congruent for the measurement of syntactic complexity and cognitive demand?

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- 20 languages with two versions of annotations were drawn from the UD 2.2 and SUD 2.2 projects to form 20 corresponding treebanks.
- Arabic (ara), Bulgarian (bul), Catalan (cat), Chinese (chi), Czech (cze), Danish (dan), Dutch (dut), Greek (ell), English (eng), Basque (eus), German (ger), Hungarian (hun), Italian (ita), Japanese (jpn), Portuguese (por), Romanian (rum), Slovenian (slv), Spanish(sp), Swedish (swe) and Turkish (tur), corresponding to Liu (2008).

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- Calculated the MDDs of all 20 treebank-pairs based on UD and SUD in accordance with formula (2) and presented with reference to Liu's (2008: 174)

- The MDD of a treebank can be defined as:

- MDD (**the treebank**) = $\frac{1}{n-s}\sum_{i=1}^{n-s}|\mathrm{DD}_i|$ (2)

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- Conducted a one-way between-subjects analysis of variance (ANOVA) test.

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- The result shows that the values of MDD changed along with the annotation schemes adopted, F (2, 57) =4.48, p = .016 < .05, η^2 = .14,
- The Tukey's post hoc indicates that no significant difference exists between MDDs based on SUD annotation scheme (M = 2.52, SD = .39) and those based on Liu (2008) (M = 2.54, SD = .48).
- Moreover, MDDs based on SUD and Liu (2008) are significantly shorter than those based on the semantic-oriented UD annotation scheme (M = 2.86, SD = .32).

3.2 Results and Discussion: Annotation Scheme and Mean Dependency Distance

- Theoretically, it is believed that annotation schemes that lead to shorter MDDs is more linguistically applicable due to that human beings tends to reduce syntactic complexity to ease the working memory burden (Osborne and Gerdes, 2019).
- The syntactic-oriented SUD is comparatively the most expedient annotation scheme to researches concerning syntactic complexity and cognitive demand when several languages are under investigation.

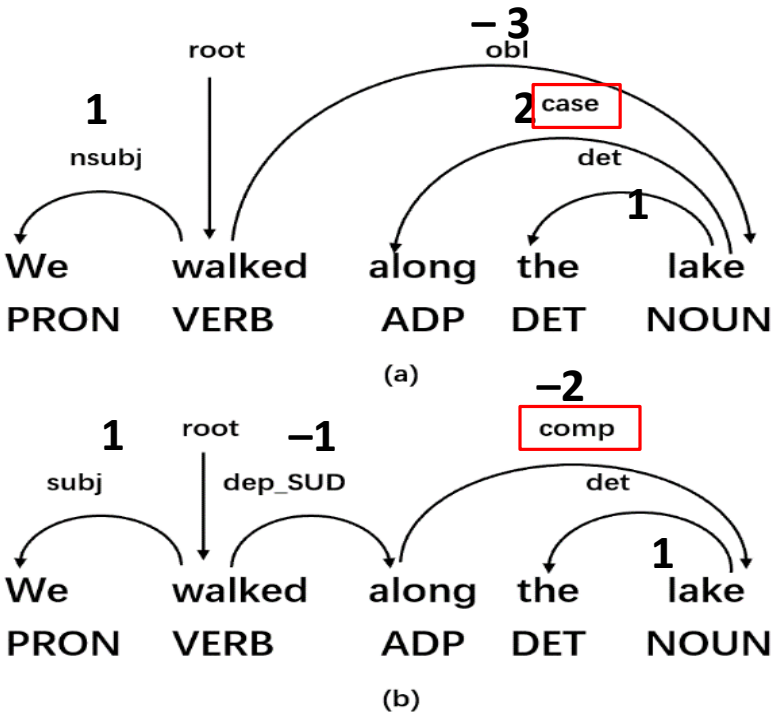
3.3 Results and Discussion: Annotation Scheme and Annotating Preference

- The Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017) in UD 2.2 and SUD 2.2 projects
- Seven genres, viz. *academic writing, biographies, fiction, interviews, news stories, travel guides* and *how-to guides,* with a total amount of 95 texts.
- **Q3:** Which dependency types account most for the distinctions between UD and SUD annotation schemes? What are the quantitative features of these dependency types?

3.3 Results and Discussion: Annotation Scheme and Annotating Preference

- The SUD annotation scheme is near-isomorphic to the UD initiative (Gerdes et al. 2018).
- The greatest difference between UD and SUD treebanks is the direction of the dependency types used to indicate the relations between function words and content words.

3.3 Results and Discussion: Annotation Scheme and Annotating Preference



- **UD** MDD:
- (|1|+|2|+|1|+|– 3|)/4=1.75.

- **SUD** MDD:
- (|1|+|– 1|+1+|– 2|)/4=1.25.

UD		SUD	
Type	**Relation**	**Type**	**Relation**
aux	auxiliary	***comp: aux***	complement: auxiliary
aux: pass	passive auxiliary	***comp: pass***	complement: passive auxiliary
cop	copula	***comp: cop***	complement: copula
mark	marker	***comp***	complement: subordinating conjunction
case	case marking		complement: adposition

Table. Detailed corresponding dependency relations in UD and SUD annotation schemes.

3.3 Results and Discussion: Annotation Scheme and Annotating Preference

- The MDDs of these 4 pairs were calculated following formula (3).

- The MDD for a specific type of dependency relation in a sample is:

- MDD (**dependency type**) = $\frac{1}{n}\sum_{i=1}^{n} \mathrm{DD}_i$ (3)

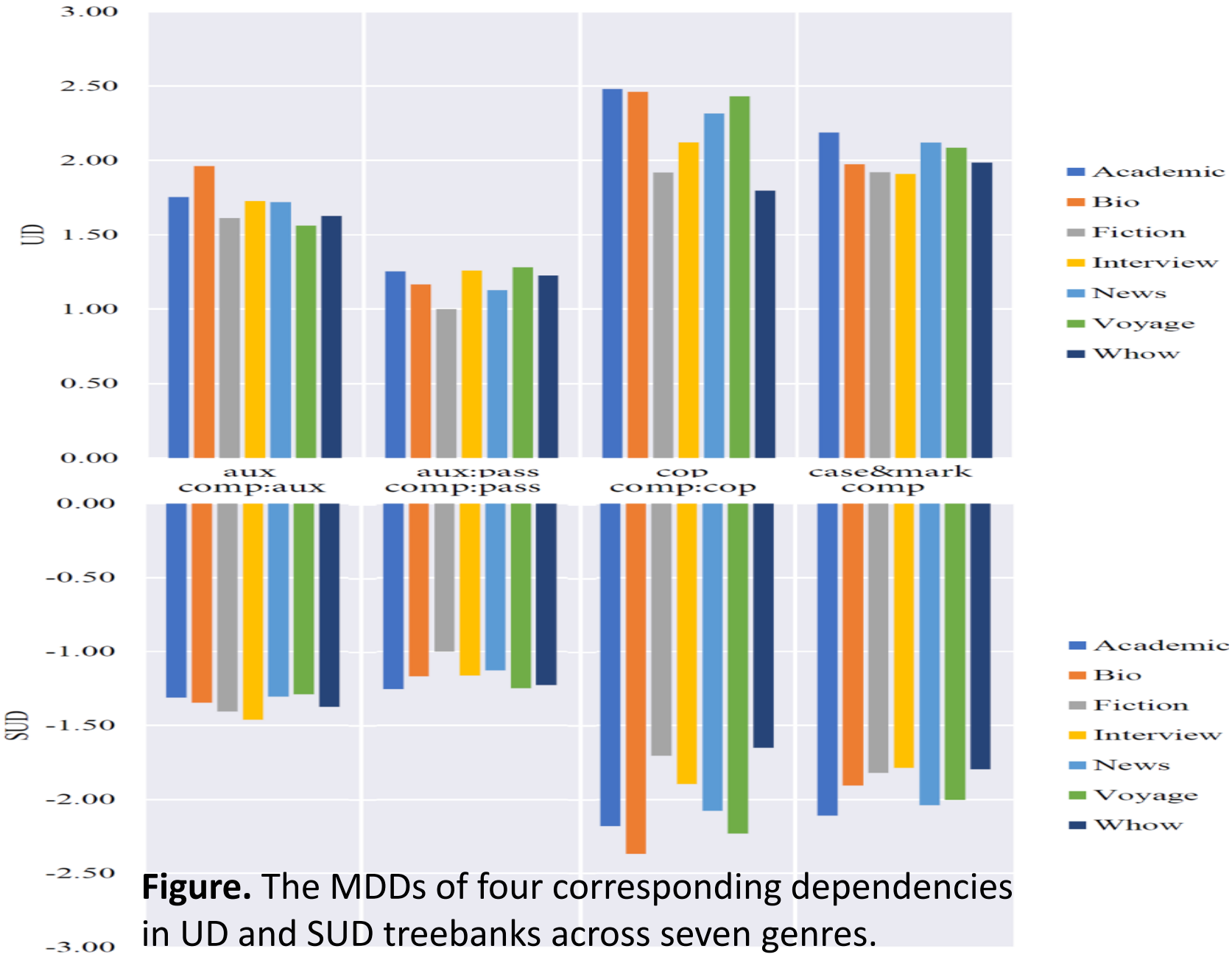


Figure. The MDDs of four corresponding dependencies in UD and SUD treebanks across seven genres.

3.3 Results and Discussion: Annotation Scheme and Annotating Preference

- The UD annotation scheme favors taking the content words as the head of function words while the SUD annotation scheme chooses the function words as heads over content words in dependency relations (Nivre, 2015; Gerdes et al., 2018; Osborne and Gerdes, 2019).
- The underlying mechanism for the distinctions between UD and SUD can be credited to the choices of head in these two annotation schemes.

4 Conclusions and Implications

- 1. The results show that, on the one hand, natural languages based on both annotation schemes follow the universal linguistic law of Dependency Distance Minimization (DDM);
- 2. On the other hand, according to the metric of Mean Dependency Distances (MDDs), the SUD annotation scheme that accords with traditional dependency syntaxes are more expedient to measure syntactic difficulty and cognitive demand.

4 Conclusions and Implications

- 3. The reason for the distinctions between UD and SUD is the dependency types indicating the relations between content words and function words. The UD annotation scheme prefers a semantic orientation, while the SUD favours a syntactic orientation which holds a function-word priority.

4 Conclusions and Implications

- Large treebanks with varieties of languages, genres or different sentence lengths are highly recommended for future researches. Meanwhile, studies on NLP and theoretical linguistics might also provide some thoughts to the questions unanswered in current study.

Thank you for your attention!

Jianwei Yan & Haitao Liu

Department of Linguistics, Zhejiang University

jwyan@zju.edu.cn & lhtzju@gmail.com