Xinying Chen, Kim Gerdes

How Do Universal Dependencies Distinguish Language Groups?

Abstract: The present paper shows how the current Universal Dependency treebanks can be used for language typology studies and can reveal structural syntactic features of languages. Two methods, one existing method and one newly proposed method, based on dependency treebanks as typological measurements, are presented and tested in order to assess both the coherence of the underlying syntactic data and the validity of the methods themselves. The results show that both methods are valid for positioning a language in the typological continuum, although they probably reveal different typological features of languages.

Keywords: treebanks, Universal Dependencies, typology, language family, Directional Dependency Distance

1 Introduction

Measuring cross-linguistic similarities and differences is one of the primary tasks of modern typology (Bickel, 2007), both theoretically and empirically. Modern language typology research (Croft 2002; Song 2001), mostly based on Greenberg (1963), focuses less on lexical similarity and relies rather on various linguistics indices for language classification, and generally puts much emphasis on the syntactic order (word order) of some grammatical relations in a sentence (Haspelmath et al. 2005).

However, a similar idea was proposed by Tesnière, before Greenberg, which attempts to classify languages by using a full analysis of a sentence based on binary grammatical relations rather than putting attention on some basic relations such as OV or VO (Liu 2010). One intuition about the difference can be that some basic grammatical relations are sufficient for detecting cross-linguistic similarities and differences already, at least for that time, the other factors being correlated with the basic distinctions of type OV/VO. Yet, several studies sug-

Xinying Chen: Xi'an Jiaotong University, chenxinying@mail.xjtu.edu.cn
Kim Gerdes: Sorbonne Nouvelle, kim@gerdes.fr

https://doi.org/10.1515/9783110573565-002

gest that this is not empirically true and it seems that combined measures on all grammatical relations, not only on the verbal and nominal arguments, provide better typological indicators than one or several specific word order measures for language classification, which may lead to conflicting conclusions (Liu 2010; Liu & Li 2010; Abramov & Mehler 2011; Liu & Xu 2011, 2012; Liu & Cong 2013). An alternative idea would be that it is just too difficult to do the research based on a complete analysis right now due to the limitations of available resources. If this is the case, then it is important to systematically replicate or test previous studies with new available data, which is also one of the motivations of this study since a suitable, newly emerged, multi-language treebank data - the Universal Dependencies treebanks - has appeared.

New available data always attracts the interest of linguists. In the case of empirical typology study, Liu (2010) used dependency treebanks of 20 languages to test and discuss some principal ideas of Tesnière and the research provided very convincing results which are in favour of Tesnière's classification. Liu (2010) completed Tesnière's attempt by developing an innovative quantitative measure of word orders, namely, the proportions of positive and negative dependencies. This measurement captures well the differences between head-initial and head-final languages. Empirical data based on 20 languages is provided along with the measurements, which makes the discussion very convincing and indicates the potential of this measurement for new data sets. However, there is a significant problem concerning this kind of Treebank based studies. The differences of annotation schemes can distort the analysis results for similar language structures or phenomena. For instance, Liu (2010) had a problem of placing Danish appropriately in his language continuum due to the peculiar annotation scheme used for the Danish dependency treebank. Addressing this problem and testing the validity and effectiveness of these measures with new available data, the Universal Dependencies treebanks, is the second motivation of the present study.

The last question we want to address is whether there are new ways of quantifying word orders for language classification and how this could be done? Following the success of dependency distance in linguistic studies (Liu et al. 2017), we propose a new way of measuring word orders, namely, directional dependency distance for this task (see section 2 for more details). The rationale is that since dependency distance reflects some universal patterns of human languages, such as dependency distance minimization (Liu et al. 2017), the details of the distance distribution in treebanks should be valuable for comparing the similarities and differences of languages. We test the proposed meas-

urement with the Universal Dependencies treebanks. Results and discussions are presented below.

The paper is structured as follows. Section 2 describes the data set, the Universal Dependencies treebanks, and introduces the new approach, directional dependency distance. Section 3 presents the empirical results and discussions of applying Liu's proportional +/– dependency approach as well as the directional dependency distance approach to the Universal Dependencies treebanks for the language classification task. Finally, Section 4 presents our conclusions.

2 Materials and Methods

Following the idea of investigating the typological similarities and differences of languages based on authentic treebank data, the present work specifically focuses on whether and how the Universal Dependencies treebank set allows us to recognize language families based on purely empirical structural data.

Universal Dependencies (UD) is a project of developing a cross-linguistically consistent tree-bank annotation scheme for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of Stanford dependencies (de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. UD is also an open resource which allows for easy replication and validation of the experiments (all Treebank data on its page is fully open and accessible to everyone). For the present paper, we used the UD_V2.0 dataset (Nivre et al., 2017) for our study since it was the most recent data when we started this work.

There are two notable advantages of using this data set for language classification studies. Firstly, it is the sheer size of the data set: It includes 70 treebanks of 50 languages, 63 of which have more than 10,000 tokens. Secondly, and most importantly, all UD treebanks use the same annotation scheme, at least in theory, as we shall see below. The few previous studies of empirical language classification based on treebank data (Liu 2010; Liu & Xu 2011, 2012) still had to rely on much fewer treebanks with heterogeneous annotation schemes. Although already relatively satisfying results were obtained, the question of identifying the source of the observed differences remains unsolved:

They could be actual structural differences between languages or simply annotation schema related differences (or even genre related differences). For instance, concerning the Danish problem mentioned in the previous section, UD can, to a certain extent, reduce the difficulty by providing a unified framework for all languages.

However, there are also drawbacks of the UD 2.0 scheme for this task. First, UD treebanks are of fundamentally different sizes. We remove the relatively sparse languages, namely treebanks with less than 10,000 tokens, from the dataset. General information about the treebanks used for this study is provided in Appendix 1 (taken from the Universal Dependencies project page, http://universaldependencies.org). Different treebanks of the same language are kept separate for consistency measures and then combined for the main classification tasks.

Secondly, some dependencies have arbitrarily been assigned to a left-to-right bouquet structure (all subsequent tokens depend on the first token). See Gerdes & Kahane (2016) for a description and for alternatives to this choice. Thereby, we only kept syntagmatic core relations and removed fixed, flat, conj, compound, and root relations from our measurements as their direction and length are universally fixed in the annotation guide and don't indicate any interesting difference between languages.

After preparing the data for the analysis, we applied the proportional approach (Liu 2010) to test the validity and effectiveness of this approach. To be specific, we computed the percentage of positive dependency relations (corresponding to a head-initial structure) versus the percentage of negative dependency relations (corresponding to a head-final structure) in a language.

Then, we applied our newly proposed approach, Directional Dependency Distance, for the same task. We define Directional Dependency Distance (DDD) as the product of the dependency distance and the direction, thus including negative values. More specifically, we computed the DDD per language by computing the difference of the node index and the governor index for each node, adding those values up and dividing by the number of links. The DDD of a language L is thus defined by its dependencies as follows:

$$\mathrm{DDD} = \frac{\sum distance(d)}{\sum frequency(d)} \tag{1}$$

To illustrate this point, we computed the Dependency Distance between the word *This* (word id 1) and the word *example* (word id 4) in the Figure 1 as: 1 – 4 = –3 (a head-final structure). The DDD of the sentence in Figure 1 is: [(1 – 4) + (2 – 4) + (3 – 4)] / 3 = –2.

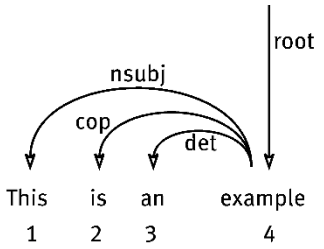

Fig. 1: A simple example of a UD tree.

We then compare and discuss the results of these two approaches. Details are presented below.

3 Results and Discussion

In this section, we present the results of applying two dependency-based word order measurements to 63 treebanks which include 43 languages.

3.1 Proportional distribution of Universal Dependencies

The percentage of positive and negative dependency distances for each language is illustrated in Figure 2 and corresponding values for treebanks are illustrated in Figure 3. For detailed values see Appendix 2.

According to Figure 2, this approach can distinguish Indo-European languages from other languages well. There are 12 languages that do not belong to the Indo-European families in our data set. Four of them, namely, Basque, Estonian, Finnish, and Hebrew (circled in Figure 5), are placed into the Indo-European continuum where they should not appear. Simply put, we can say that the accuracy of distinguishing Indo-European languages of this approach is (43 – 4) * 100 / 43 = 90.7%, which seems like a rather good result. However, if we take a closer look at the Indo-European continuum, we can see that this approach performs poorly for distinguishing sub-groups of Indo-European families. For instance, Spanish (Romance language), Norwegian (Germanic language), and Czech (Slavic language) are placed next to each other. Even worse, Croatian (Slavic language), Portuguese (Romance language), Latin (Latin language), Finnish (Finnic language, does not belong to the Indo-European group), Ancient Greek (Greek language), Slovak (Slavic language), Danish (Germanic

language), Basque (Basque language, does not belong to the Indo-European group), Estonian (Finnic language, does not belong to the Indo-European group), Persian (Iranian language), and Hebrew (Semitic language, does not belong to the Indo-European group) are next to each other. Our results are not as satisfying as the results with 20 languages reported by Liu (2010), which rather successfully distinguished sub-groups of Indo-European families. The cause of this difference is hard to identify. Possible explanations can be: 1) the genre differences; 2) the treebank size differences; 3) the approach is good at coarse distinction but it is not decent for more fine-grained task; 4) the inhomogeneous annotation schemes are better suited for difference measures than the rather homogenous UD annotation scheme (Different from common annotations, UD promotes content words, instead of function words, as the heads of dependencies. This strategy may relatively weaken the differences between languages). If we had access to comparable treebanks – or even better to parallel treebanks – which are still a very scarce resource, available for only very few languages at this stage, we decide to leave this question to future research.

Figure 3 reveals some incoherence of the current state of the UD. It indicates the different places taken by the English (arrows on the left side) and the French (arrows on the right side) treebanks of UD_V2.0. Similar situations arise for Italian, Czech, Russian, etc. We obtain no satisfying results for any multi-treebank language dataset. Note also that the parallel treebanks from the ParTUT team (the Multilingual Turin University Parallel Treebank, Sanguinetti 2013, circled in Figure 3) coherently have a lower percentage of positive dependency links than their counterparts for English, French, and Italian. It seems that any derived typological classification could remain quite treebank dependent at the current state of UD.

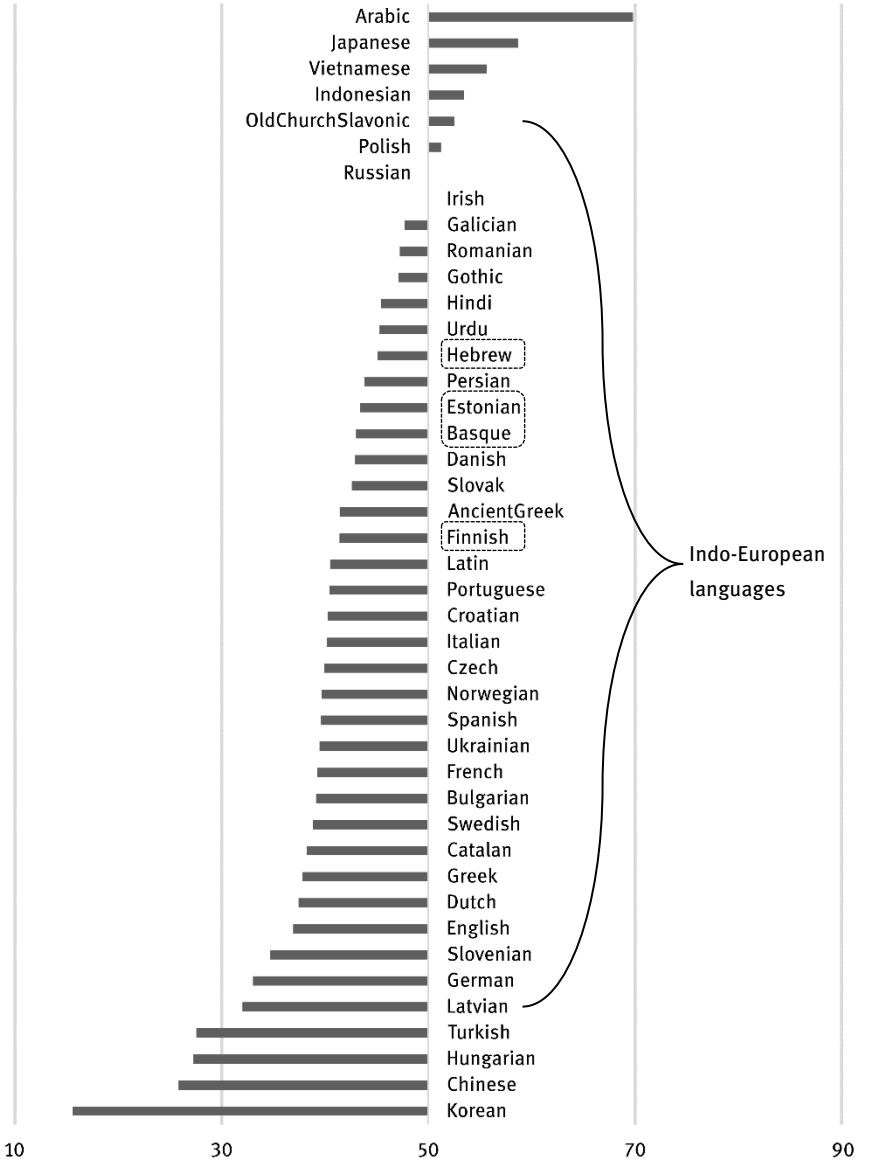

Fig. 2: Languages ordered by % of positive links

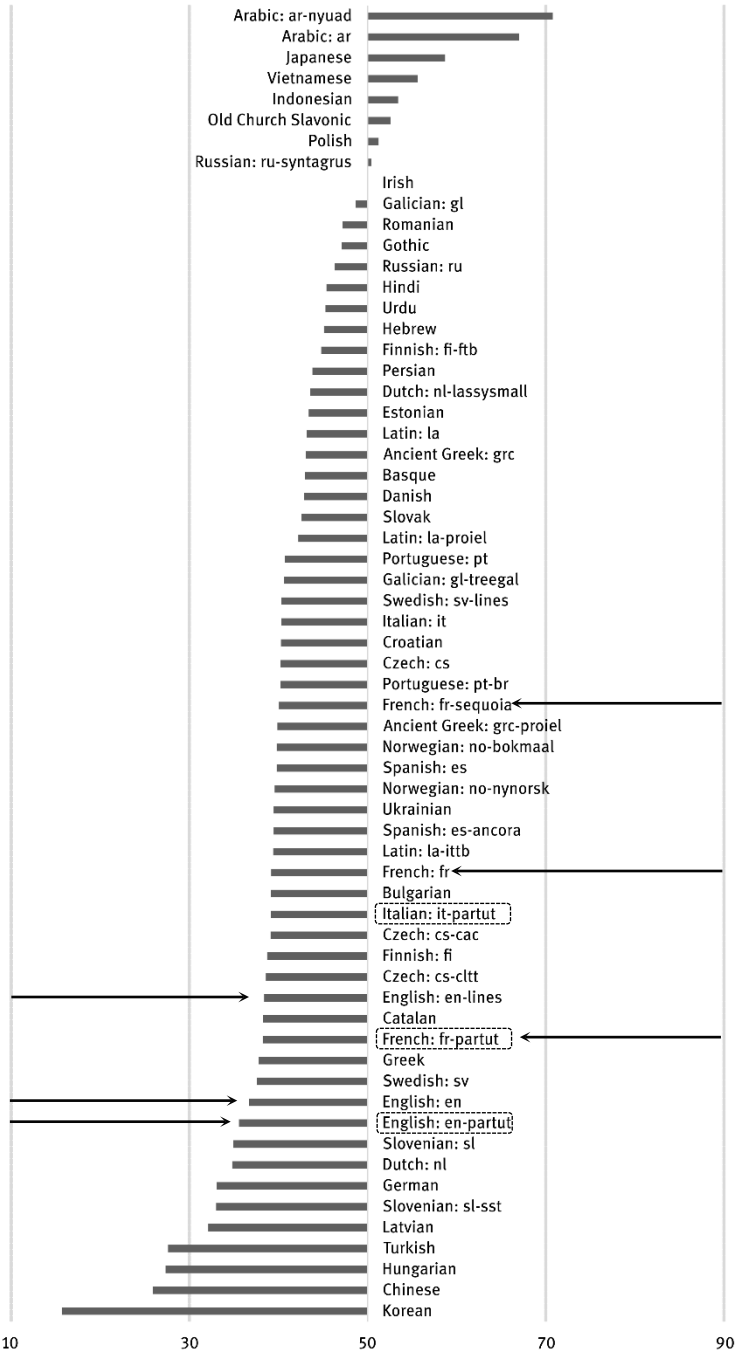

Fig. 3: Treebanks ordered by % of positive links

3.2 DDD of languages and treebanks

The DDD for each language is illustrated in Figure 4 and corresponding values for treebanks are illustrated in Figure 5. For detailed values see Appendix 3.

Different from the results of Figure 2, DDD has a relatively lower accuracy for distinguishing Indo-European languages from other languages. Although the number of languages that manage to make their way into the Indo-European continuum is 4, which is the same with the results of the simple measure of Figure 2, there are also Indo-European languages, such as Hindi and Urdu, mixed into the other languages. Yet, in general, DDD still does a good job for this coarse distinction task with an accuracy (43 – 7) * 100 / 43 = 83.7%. Besides the broad sense accuracy, some interesting details are worthy of more attention. Observe how Japanese finds its natural position close to Korean, Turkish, and Hungarian (they are all agglutinative languages) in the DDD measure, whereas the direction percentage measure places Japanese right next to Arabic. One should bear in mind that the simple accuracy may not be a decent assessment for deciding which approach is 'better'.

On the other hand, DDD does deal with sub-groups of Indo-European families better. Although the Germanic language group spreads across the spectrum, the Romance languages, however, are very well clustered around an average distance of about 0.8.

Similar to Figure 3, Figure 5 also reveals some incoherence of the current state of the UD but with different patterns. It indicates that the different places taken by the English (arrows on the left side) treebanks are more distant than that of the French (arrows on the right side), which is opposite to Figure 3. Although the absolute values are not as extremely different as the position suggests (en: 0.36, 0.53, 0.77; fr: 0.64, 0.76, 0.79), it confirms that any derived typological classification seems to remain quite Treebank-dependent at the current state of UD. Another detail is also coherent with what we find in Figure 3. The parallel treebanks from the ParTUT team (circled in Figure 5) show a similar pattern. They coherently have lower dependency directions than their counterparts for English, French, and Italian. It is tempting to attribute this difference to differences in the guidelines used by different teams in the annotation process. So maybe the difference is rather due to the syntactic structure of "Translationese" (Baroni & Bernardini 2005), that has shorter dependency links for the mostly head-initial languages included in ParTUT. More generally, this shows how these methods also allow for detecting common ground and outliers in the process of treebank development, making them a tool of choice for error-mining a treebank.

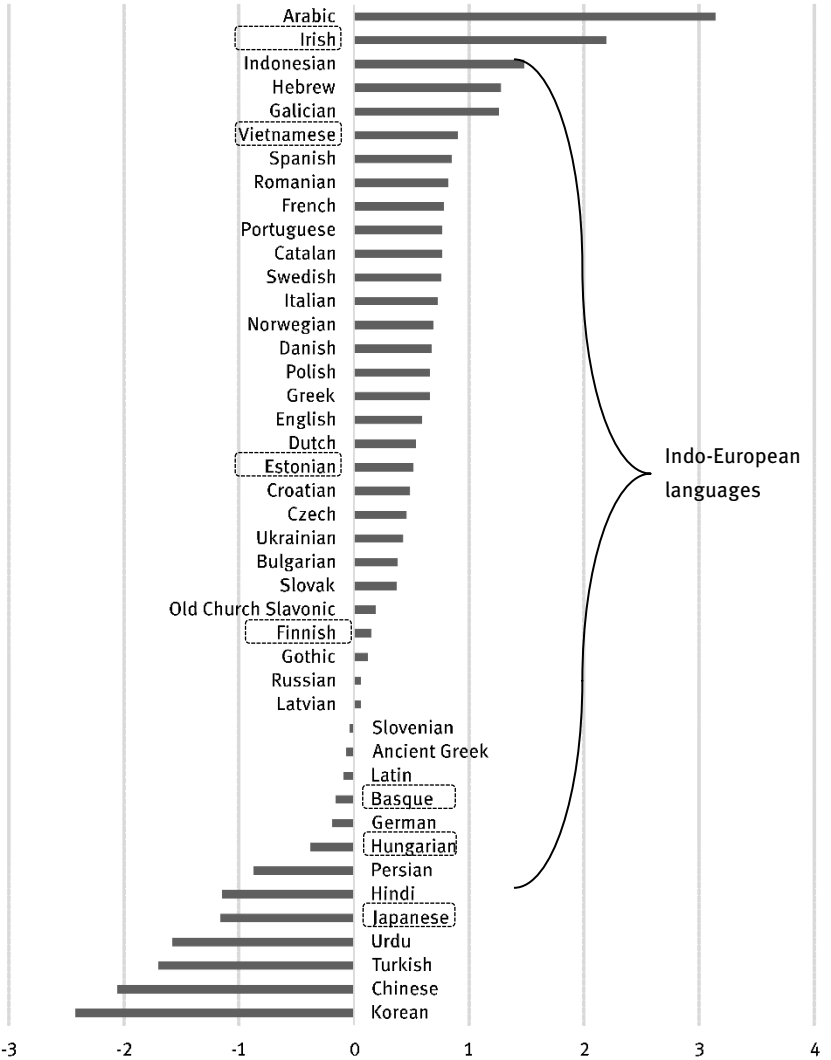

Fig. 4: Languages ordered by directional dependency distance

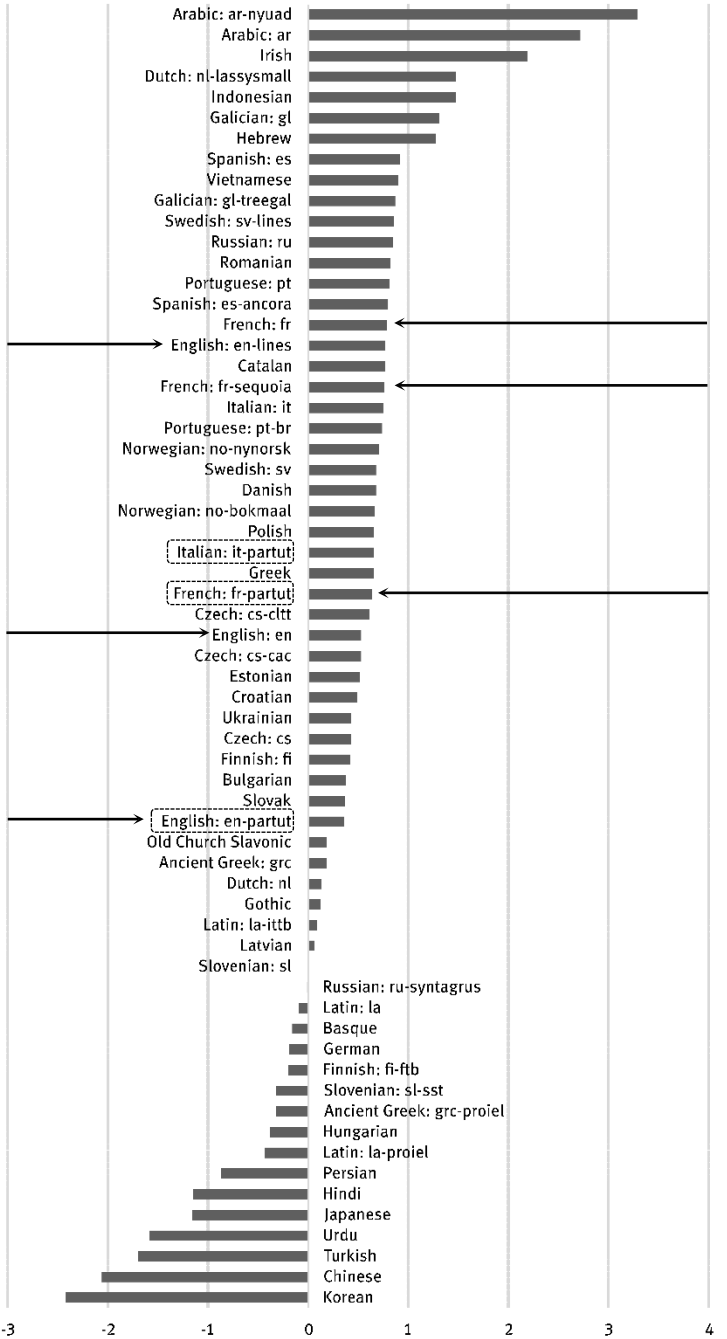

Fig. 5: Treebanks ordered by directional dependency distance

4 Conclusion

Our research shows that Tesnière's idea, which is classifying languages by head-initial and head-final features of all grammatical structures in a sentence, is compatible with empirical analysis.

Liu (2010) completes this idea by measures of the proportions of positive and negative dependency of a language. The measure proposed alongside empirical support of 20 languages is tested in this study with a new dataset, the UD_V2.0 dataset. Our results show that this approach performs very well in a coarse sense but not as satisfying as previous studies (Liu 2010) suggest. Although, the cause of less satisfying results cannot be identified in the current state, we still could cautiously conclude that this approach is proved to be valid.

We also propose and present a new measurement in this study. Results show that this DDD approach performs better in more fine-grained task. It is possible that DDD reflects different typological features in comparison to previous measurement. In general, it is also an effective approach for distinguishing language groups.

Nevertheless, one should not ignore language classification studies that are less 'orthodox', which carry more engineering genres rather than linguistic ones. Many studies seem to obtain more satisfying results with multi parameter data or multiple metrics that provide more details than simple percentages or mean distances (Chen & Gerdes 2017; Liu & Li 2010; Abramov & Mehler 2011; Liu & Xu 2011, 2012; Liu & Cong 2013). However, it remains a job for linguists in the future to improve existing approaches of analysing language data and provide comprehensive explanations for all these phenomena.

Acknowledgement: This work is supported by the Fundamental Research Funds for the Central Universities (Program of Chinese Function Words in Syntactic Networks, Xi'an Jiaotong University), Social Science Fund of Shaanxi Province (2015K001).

References

Abramov, Olga & Alexander Mehler. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics, 18*(4), 291–336.

Baroni, Marco & Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing, 21*(3), 259–274.

Bickel, Balthasar. 2007. Typology in the 21st century: major current developments. *Linguistic Typology, 11*(1), 239–251.

Chen, Xinying & Kim Gerdes. 2017, September. Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy (No. 139, 54–63). Linköping University Electronic Press.

Croft, William. 2002. *Typology and Universals*. Cambridge: Cambridge University Press.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014, May. Universal Stanford dependencies: A cross-linguistic typology. In LREC (Vol. 14, 4585–92).

Gerdes, Kim, & Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In LAW X (2016) The 10th Linguistic Annotation Workshop: 131.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (Ed.), *Universals of Language* (pp.73–113). Cambridge, MA: MIT Press.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

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

Liu, Haitao & Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin, 58*(10), 1139–1144.

Liu, Haitao & Wenwen Li. 2010. Language clusters based on linguistic complex networks. *Chinese Science Bulletin, 55*(30), 3458–3465.

Liu, Haitao & Chunshan Xu. 2011. Can syntactic networks indicate morphological complexity of a language? *EPL (Europhysics Letters), 93*(2), 28005.

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

Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, 171–193.

Nivre, Joakim, & Lars Ahrenberg Željko Agić. 2017. Universal Dependencies 2.0 CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2011. A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

Sanguinetti, Manuela, Cristina Bosco, & Leonardo Lesmo. 2013. Dependency and Constituency in Translation Shift Analysis. In Proceedings of the 2nd Conference on Dependency Linguistics (DepLing 2013), 282–291.

Song, Jae Jung. 2014. *Linguistic Typology: Morphology and Syntax*. New York: Routledge.

Zeman, Daniel. 2008, May. Reusable Tagset Conversion Using Tagset Drivers. In LREC, 28–30.

Appendix

Tab. 1: General information of treebanks used in the study is taken from the Universal Dependencies project page, http://universaldependencies.org.

Language	Numbers of Treebanks	Size in Tokens (k)	Language Family
Ancient Greek	2	211; 202	Indo-European, Greek
Arabic	2	738; 282	Afro-Asiatic, Semitic
Basque	1	121	Basque
Bulgarian	1	156	Indo-European, Slavic
Catalan	1	531	Indo-European, Romance
Chinese	1	123	Sino-Tibetan
Croatian	1	197	Indo-European, Slavic
Czech	3	1,506; 494; 35	Indo-European, Slavic
Danish	1	100	Indo-European, Germanic
Dutch	2	208; 101	Indo-European, Germanic
English	3	254; 82; 49	Indo-European, Germanic
Estonian	1	106	Uralic, Finnic
Finnish	2	202; 159	Uralic, Finnic
French	3	402; 70; 28	Indo-European, Romance
Galician	2	138; 25	Indo-European, Romance
German	1	292	Indo-European, Germanic
Gothic	1	55	Indo-European, Germanic
Greek	1	63	Indo-European, Greek
Hebrew	1	161	Afro-Asiatic, Semitic
Hindi	1	351	Indo-European, Indic
Hungarian	1	42	Uralic, Ugric
Indonesian	1	121	Austronesian
Irish	1	23	Indo-European, Celtic
Italian	2	293; 55	Indo-European, Romance
Japanese	1	186	Japanese
Korean	1	74	Korean
Latin	3	291; 171; 29	Indo-European, Latin
Latvian	1	90	Indo-European, Baltic
Norwegian	2	310; 301	Indo-European, Germanic
Old Church Slavonic	1	57	Indo-European, Slavic

Language	Numbers of Treebanks	Size in Tokens (k)	Language Family
Persian	1	152	Indo-European, Iranian
Polish	1	83	Indo-European, Slavic
Portuguese	2	319; 227	Indo-European, Romance
Romanian	1	21	Indo-European, Romance
Russian	2	99; 1,107	Indo-European, Slavic
Slovak	1	106	Indo-European, Slavic
Slovenian	2	140; 29	Indo-European, Slavic
Spanish	2	549; 431	Indo-European, Romance
Swedish	2	96; 79	Indo-European, Germanic
Turkish	1	58	Turkic, Southwestern
Ukrainian	1	100	Indo-European, Slavic
Urdu	1	138	Indo-European, Indic
Vietnamese	1	43	Austro-Asiatic

Tab. 2: Percentage of positive links for languages and treebanks

Language	% of positive links	Language	% of positive links
Ancient Greek (all)	41.41	Indonesian	53.44
grc	43.04	Irish	49.93
grc-proiel	39.88	Italian (all)	40.17
Arabic (all)	69.81	it	40.33
ar	67.01	it-partut	39.13
ar-nyuad	70.79	Japanese	58.68
Basque	43.00	Korean	15.67
Bulgarian	39.14	Latin (all)	40.47
Catalan	38.26	la	43.16
Chinese	25.89	la-ittb	39.39
Croatian	40.27	la-proiel	42.21
Czech (all)	39.90	Latvian	32.10
cs	40.21	Norwegian (all)	39.67
cs-cac	39.11	no-bokmaal	39.79
cs-cltt	38.59	no-nynorsk	39.55
Danish	42.89	Old Church Slavonic	52.55
Dutch (all)	37.44	Persian	43.81

Language	% of positive links	Language	% of positive links
nl	34.81	Polish	51.23
nl-lassysmall	43.55	Portuguese (all)	40.41
English (all)	36.91	pt	40.70
en	36.70	pt-br	40.20
en-lines	38.38	Romanian	47.22
en-partut	35.58	Russian (all)	50.08
Estonian	43.39	ru	46.31
Finnish (all)	41.37	ru-syntagrus	50.42
fi	38.72	Slovak	42.58
fi-ftb	44.78	Slovenian (all)	34.68
French (all)	39.23	sl	34.92
fr	39.16	sl-sst	33.00
fr-partut	38.24	Spanish (all)	39.60
fr-sequoia	40.05	es	39.78
Galician (all)	47.72	es-ancora	39.44
gl	48.64	Swedish (all)	38.85
gl-treegal	40.64	sv	37.58
German	33.04	sv-lines	40.33
Gothic	47.11	Turkish	27.61
Greek	37.79	Ukrainian	39.45
Hebrew	45.09	Urdu	45.26
Hindi	45.39	Vietnamese	55.65
Hungarian	27.32		

Tab. 3: DDD for languages and treebanks

Language	DDD	Language	DDD
Ancient Greek (all)	−0.07	Indonesian	1.48
grc	0.19	Irish	2.19
grc-proiel	−0.32	Italian (all)	0.73
Arabic (all)	3.14	it	0.75
ar	2.72	it-partut	0.66
ar-nyuad	3.29	Japanese	−1.16
Basque	−0.16	Korean	−2.42
Bulgarian	0.38	Latin (all)	−0.09

Language	DDD	Language	DDD
Catalan	0.77	la	−0.09
Chinese	−2.06	la-ittb	0.09
Croatian	0.49	la-proiel	−0.43
Czech (all)	0.46	Latvian	0.06
cs	0.43	Norwegian (all)	0.69
cs-cac	0.53	no-bokmaal	0.67
cs-cltt	0.61	no-nynorsk	0.71
Danish	0.68	Old Church Slavonic	0.19
Dutch (all)	0.54	Persian	−0.87
nl	0.13	Polish	0.66
nl-lassysmall	1.48	Portuguese (all)	0.77
English (all)	0.59	pt	0.81
en	0.53	pt-br	0.74
en-lines	0.77	Romanian	0.82
en-partut	0.36	Russian (all)	0.06
Estonian	0.52	ru	0.85
Finnish (all)	0.15	ru-syntagrus	−0.01
fi	0.42	Slovak	0.37
fi-ftb	−0.20	Slovenian (all)	−0.04
French (all)	0.78	sl	0.00
fr	0.79	sl-sst	−0.32
fr-partut	0.64	Spanish (all)	0.85
fr-sequoia	0.76	es	0.92
Galician (all)	1.26	es-ancora	0.80
gl	1.31	Swedish (all)	0.76
gl-treegal	0.87	sv	0.68
German	−0.19	sv-lines	0.86
Gothic	0.12	Turkish	−1.70
Greek	0.66	Ukrainian	0.43
Hebrew	1.28	Urdu	−1.58
Hindi	−1.15	Vietnamese	0.90
Hungarian	−0.38		