Curve-fitting of frequency distributions of dependency distances in a multi-lingual parallel corpus

Masanori Oya
School of Global Japanese Studies, Meiji University,
4-21-1, Nakano, Tokyo, Japan
masanori_oya2019@meiji.ac.jp

Abstract

This study discusses the curve-fitting results of the frequency distributions of dependency distances in sentences from a multi-lingual parallel corpus. It assumes that these distributions fit a mathematically well-defined distribution (the right truncated modified Zipf-Alekseev distribution) quite well, which indicates that these distributions of dependency distances reflect an aspect of the universal properties of natural language. However, the results of the curve-fitting of frequency distributions of the dependency distances of different dependency types do not demonstrate a suitable fit to the right truncated modified Zipf-Alekseev distribution, suggesting the necessity of further research.

1 Introduction

Dependency distance has been the center of focus of research on memory burden and syntactic complexity (Gibson, 1998, 2000; Gildea & Temperley, 2010; Grodner & Gibson, 2005; Li & Yan, 2021; Liu, 2007, 2008; Liu et al., 2017; Oya, 2013, 2021). The dependency distance between words in a dependency relation can be easily calculated; for example, in the sentence "Sarah has written an article in two months," the noun *Sarah* depends on the auxiliary verb *has* as the subject, and the dependency distance between them is one. The noun *article* depends on *written* as its object, and the dependency distance between them is two.

Dependency distance has been argued as an aspect of the universal properties of natural languages; more specifically, shorter dependency distances are preferred to longer ones, possibly due to the upper bound of the short-term memory of humans (Gibson, 2000). It has been found out that the threshold of dependency distance is four across different natural languages (Liu, 2008; Oya, 2021). This means that the frequencies of dependency distances one, two, or three are much higher than the frequencies of dependency distances four or larger. When we plot the frequency distributions of dependency distances on an x-y plane with the x axis being the dependency distance and the y axis its frequency, then the graph has a long tail to the right.

In this context, frequency distributions of dependency distances have been discovered to fit the right truncated modified Zipf-Alekseev distribution (henceforth ZA distribution) quite well (Jiang & Liu, 2015; Liu, 2009; Ouyang & Jiang, 2017). The ZA distribution formula is illustrated below (Ouyang & Jiang, 2017; Popescu et al., 2014; Li and Yan, 2021):

$$y = cx^{a+b\ln x}$$

In the formula above, x is a dependency distance, y is its frequency, *c* is a constant, and the parameters *a* and *b* vary along with the fitness of the distributions of the frequencies. According to Li and Yan (2021), when the ZA distribution was fitted to the frequency distribution of dependency distances across essays written by Japanese EFL

learners of different proficiencies, the results of good fitness were obtained.

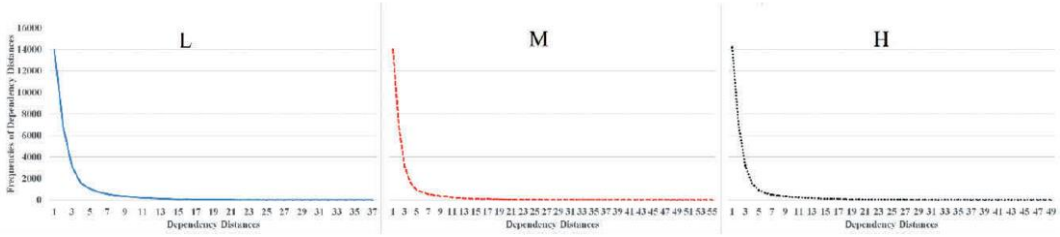

Figure 1. Frequency distributions of the dependency distances of Lower, Middle and High groups (Li & Yan, 2021, p.180)

Li and Yan (2021) also found out that learner proficiencies are reflected in the different parameters of the ZA distribution. That is, parameter *a* increases, and parameter *b* decreases, from the lower (L in the table below), the intermediate (M), and to the higher proficiency learner groups (H).

Group	a	b	n	α	X^2	$P(X^2)$	DF	C	R^2
L	0.66	0.4634	45	0.4646	366.69	0	40	0.0112	0.9969
M	0.6982	0.4056	85	0.4605	611.47	0	60	0.0201	0.9944
H	0.7206	0.3898	63	0.4608	805	0	55	0.0261	0.9929

Table 1. Fitting ZA distribution to the frequency distributions of dependency distances of different groups (Li & Yan, 2021, p.181)

2. Background of this study

This study follows Li and Yan's (2021) research, attempts to fit the ZA distribution to the frequency distribution of dependency distances in sentences taken from a multi-lingual parallel corpus, and hopes to find aspects of properties that are universal across different languages. More specifically, we focus on the similarities/differences of parameters *a* and *b* across different languages in a multi-lingual parallel corpus, with the assumption that such comparisons can indicate the similarities/differences of these languages mathematically, with reference to the ZA distribution. If the parameters *a* and *b* of one language are found to be close to those of another language, it will indicate their similarity as far as their fitness to the ZA distribution is concerned. If, on the other hand, these parameters are widely different across the two languages, it will be indicative of their difference in terms of the ZA distribution.

Along with this distribution-fitting of individual languages, we also focus on the distributions of dependency distances of different dependency types, e.g., the distribution of dependency distances between a verb and its subject, and its attempt to fit to the ZA distribution. Similar to the curve-fitting of the distributions of dependency distances across all dependency types, we focus on the similarities/differences of parameters *a* and *b* of different languages in the same multi-lingual parallel corpus. This fitting is expected to reveal the behaviors of the dependency distances of different types, and to open up the possibility of the investigation of a certain aspect of the universal properties of natural languages from a more fine-grained perspective.

This study poses the following research questions:

1. Will the frequency distribution of the dependency distances in a multi-lingual parallel corpus fit suitably to the ZA distribution?
2. Will similarities/differences of languages be reflected on the parameters of the ZA distribution?
3. Will the frequency distribution of the dependency distances of different dependency types be reflected in the parameters of ZA distribution?

2.1 Data

This study uses *Parallel Universal Dependencies Treebanks 2.7* (henceforth PUD) to answer the abovementioned research questions. These treebanks were created for the purpose of the shared task on Multilingual Parsing from Raw Text to Universal Dependencies at CoNLL 2017 (http://universaldependencies.org/conll17/). PUD is a parallel corpus consisting of 20,000 sentences with aligned translation pairs across 20 languages. (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Icelandic, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish). Of the 1,000 sentences of each language, 750 were those which had been translated directly from English texts, and the remaining 250 were translated from German, French, Italian, or Spanish; these sentences had been first translated into English, and then further translated into each of these languages. The translation of these sentences was conducted manually by professional translators, and these sentences were further annotated with morphological and syntactic tags by Google. These annotations are converted into UD (Universal Dependencies) by UD community members, according to UD Ver. 2 guidelines. The UD website elucidates further details (https://universaldependencies.org/).

2.2 Procedure

We can calculate the dependency distance between two words in dependency relationship. We can obtain the dependency distance of each dependency relation in PUD through the dependency tag on every word in a sentence. It is conducted by an original Ruby script which was the same used in Oya (2021). On the basis of the output of the script, we obtained the frequencies of the dependency distances of all the dependencies in the sentences in each of the languages in PUD, as well as the frequencies of dependency distances of different dependency types. This study focuses on the frequencies of the dependency distances of the dependency types *nsubj* (dependency between a verb and its nominal subject), *obj* (dependency between a transitive verb and its direct object), *amod* (dependency between a noun and an adjective modifying the noun), and *advmod* (dependency between an adverb and a noun modified by the adverb). Then, the frequency distributions of dependency distances are calculated, and fitted to the ZA distribution by means of the Altmann-fitter v.3.1.0 (http://www.ram-verlag.biz/altmann-fitter/), which is the same application that Lin and Yan (2021) used to curve-fit their data.

2.3 Results

The table below illustrates the results of the curve-fitting of the dependency distances' frequency distributions in the 20 PUD languages to the ZA distribution. The R^2 values of all these languages (except for French) are above 0.98, which indicates that they all fit the ZA distribution quite well.

	a	*b*	n	α	X^2	$P(X^2)$	C	DF	R^2
Arabic	0.9274	0.2634	51	0.5214	244.27	0	0.0118	46	0.9985
Czech	0.3531	0.4618	45	0.4061	211.4	0	0.0114	40	0.9971
German	0.2154	0.4234	51	0.3583	411.76	0	0.0193	46	0.9925
English	0.206	0.5165	56	0.3638	490.23	0	0.0232	48	0.9924
Spanish	0.4536	0.4437	59	0.407	995.06	0	0.0427	52	0.9853
Finnish	0.3764	0.4738	44	0.4245	96.4	0	0.0061	39	0.9998
French	0.1188	0.5258	52	0.4063	1491.42	0	0.0603	47	0.9787
Hindi	0.5507	0.2748	53	0.4605	1050.47	0	0.0441	48	0.9915
Indonesian	0.6206	0.3706	46	0.4825	224.7	0	0.0116	41	0.9976
Icelandic	0.3406	0.4987	49	0.4317	290.77	0	0.0154	42	0.9957
Italian	0.2291	0.5015	63	0.4047	1263.5	0	0.0532	52	0.9816
Japanese	1.4917	0.1277	72	0.4327	1486.15	0	0.0516	67	0.9894
Korean	0.8252	0.2137	46	0.5528	527.16	0	0.0318	41	0.9967
Polish	0.5594	0.3939	42	0.4628	194.13	0	0.0106	37	0.9976
Portuguese	0.2558	0.4947	61	0.4058	1097.69	0	0.0469	51	0.9836
Russian	0.5218	0.4262	46	0.4259	384.96	0	0.0199	41	0.9936
Swedish	0.1053	0.5602	50	0.3888	537.5	0	0.0282	42	0.9903
Thai	0.5959	0.4327	42	0.5327	101.95	0	0.0046	37	0.9994
Turkish	0.7389	0.2438	40	0.5198	552.58	0	0.0327	35	0.9954
Chinese	0.7587	0.2628	47	0.4085	181.6	0	0.0085	42	0.9995

Table 2. The results of the curve-fitting of the frequency distributions of the dependency distances in the 20 languages in the PUD, to the ZA distribution.

The figure below illustrates that, when these languages are plotted according to their parameters *a* and *b*, they demonstrate a linear distribution; $R^2 = .82$, $p < .01$. In this linear distribution, we can notice that languages of the Indo-European family seem to form one cluster in which the parameter *a* is less than 0.6 and the parameter *b* is within the range of 0.4 and 0.6 (except for Hindi), while non-Indo-European languages form another cluster in which the parameter *a* lies within the range of 0.6 and 1 and the parameter *b* is less than 0.4 (except for Finnish and Thai). Japanese stands out from these two clusters.

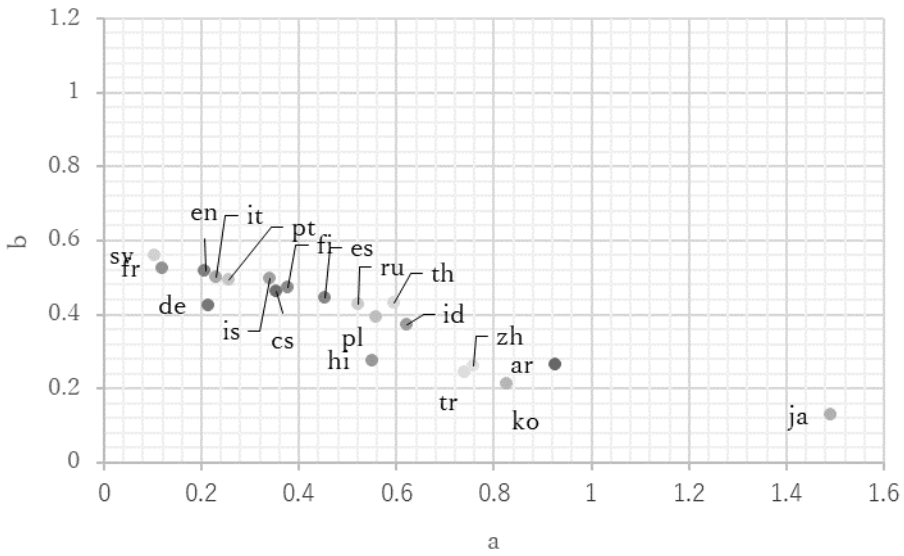

Figure 3. The plots of parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of all dependency types in the PUD are fitted to the ZA distribution

Unlike the case of the plot above which takes all the dependency types into consideration, when the dependency distances of *nsubj* in the PUD are fitted to the ZA distribution, there is a weak linear distribution of the languages in the PUD in terms of their *a* and *b* parameters of the ZA distribution; $R^2 = .27$, $p < .02$.

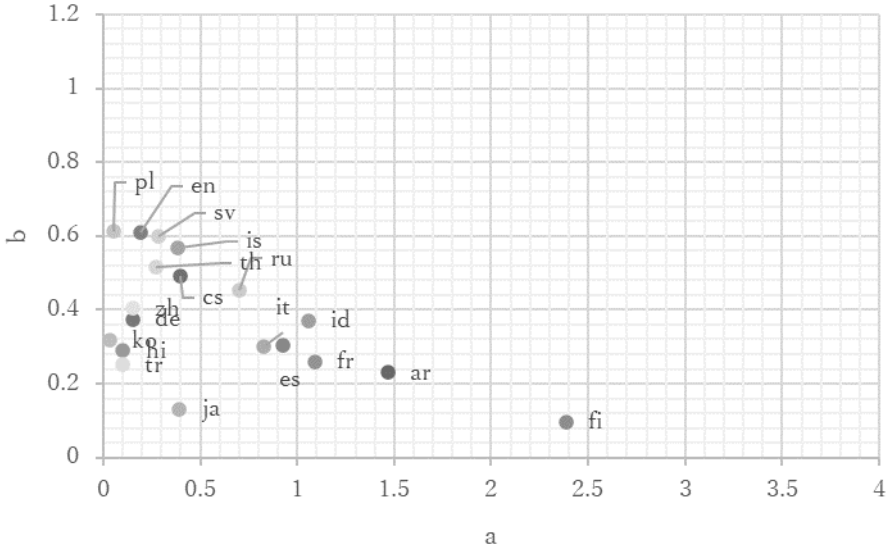

Figure 4. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *nsubj* in the PUD are fitted to the ZA distribution

When the dependency distances of *obj* in the PUD are fitted to the ZA distribution, there is no linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; $R^2 = .13$, $p < .12$.

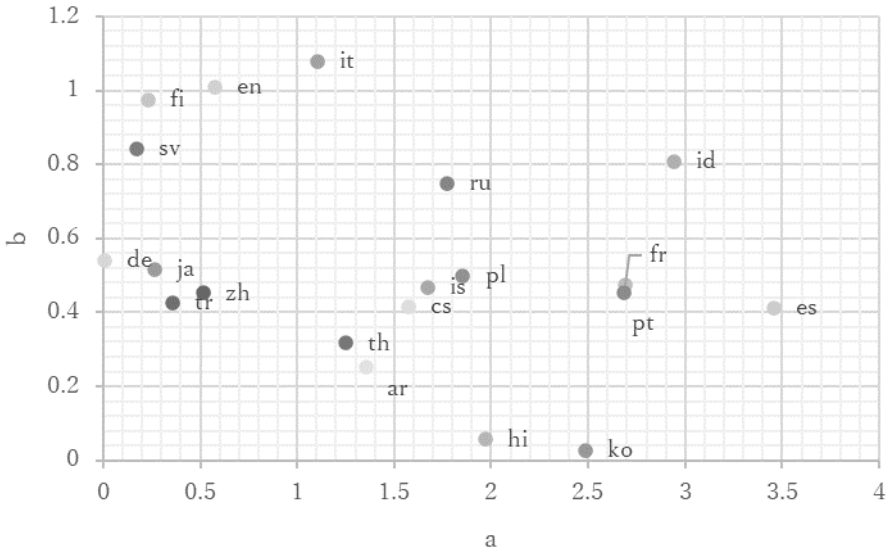

Figure 5. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *obj* in the PUD are fitted to the ZA distribution

When the dependency distances of *amod* in the PUD are fitted to the ZA distribution, there is no linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; R^2 = .08, p < .21. This non-linear distribution is partly due to the fact that the frequencies of *amod* in Arabic, Indonesian and Japanese do not fit the ZA distribution. When these frequencies are excluded, the distribution turns out to be linear; R^2 = .67, p < .01.

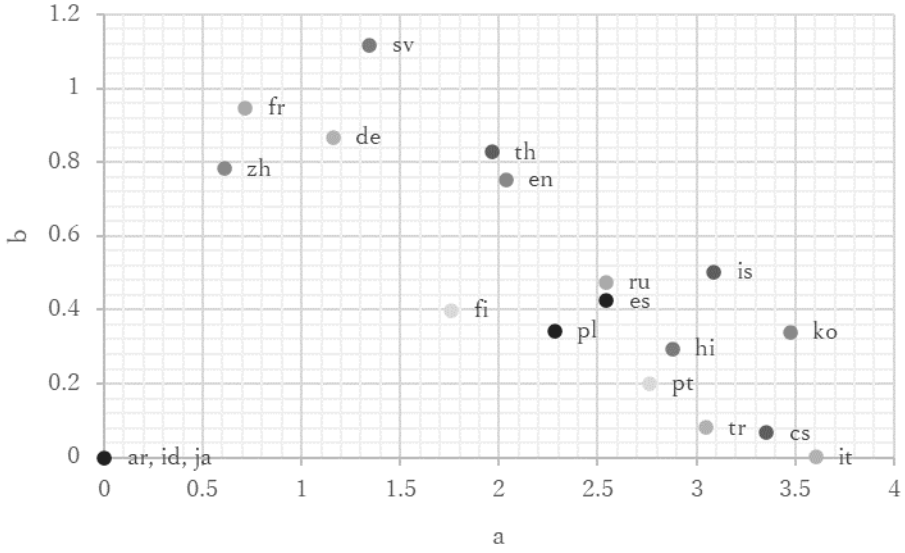

Figure 6. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *amod* in the PUD are fitted to the ZA distribution

When the dependency distances of *advmod* in the PUD are fitted to the ZA distribution, there is a weak linear distribution of the languages in the PUD in terms of their parameters *a* and *b* of the ZA distribution; R^2 = .46, p < .01.

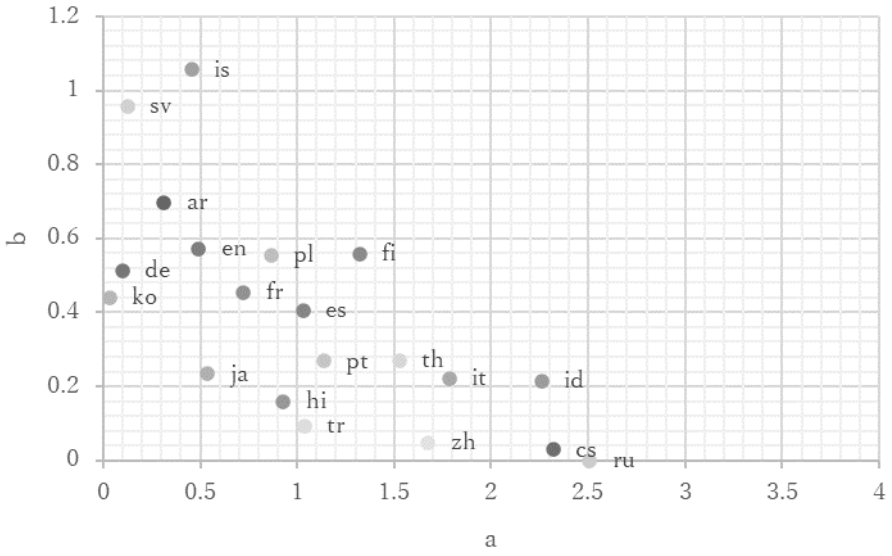

Figure 7. The plots of the parameters *a* and *b* of the languages in the PUD, when the frequency distributions of the dependency distances of *advmod* in the PUD are fitted to the ZA distribution

3. Discussion

The results answer R.Q.1 and R.Q.2, and indicate that the frequency distribution of dependency distances in a multi-lingual parallel corpus fit well to the ZA distribution (R.Q.1), and that differences between languages are reflected in the different parameters of the ZA distribution (R.Q. 2). The clusters of parameters seem to reflect the differences of language families they belong to; however, further investigations that include more languages which have not been included in PUD are necessary.

The aforementioned results do not seem to answer R.Q.3 positively. The frequency distributions of the dependency distances of different dependency types in the multi-lingual parallel corpus do not necessarily fit well with the ZA distribution. These results may be due to the fact that the same dependency type (e.g., *obj*) can be used differently in different languages, even in a parallel corpus in which sentences of different languages are aligned according to their semantic parallelism; for example, the direct object of a verb in one sentence of a language can be expressed not as the direct object of the sentence's translation in another language, and vice versa. The same is true for other dependency types. This may result in more random distribution of dependency distances, and less linear distribution of the parameters *a* and *b*.

We can notice different degrees of linear distributions of the parameters *a* and *b* across different dependency types. For example, the parameters *a* and *b* of the dependency type *nsubj* seems to have relatively higher degree of linear distribution than other dependency types. This may reflect the fact that what is expressed as the subject of a verb in one language is often expressed as the subject of a verb in other languages, and also their dependency distances (and their frequencies) are similar with each other. As such, we still need further investigation into different distributions of dependency distances of different dependency types across different languages, and finding language-(in)dependent patterns across these differences remains to be one of the goals in future research.

4. Conclusion

This study reported the results of the curve-fitting of frequency distributions of dependency distances in sentences within a multi-lingual parallel corpus. It was found that these distributions fit the ZA distribution fairly well. This indicates that these

distributions of dependency distances can reflect a certain aspect of the universal properties of natural language. However, the results of the curve-fitting of frequency distributions of the dependency distances of different dependency types do not demonstrate a suitable fit to the ZA distribution. Thus, these results require us to further investigate the behaviors of different types of dependencies in terms of their distances and their frequencies, both within the same parallel corpus and other types of corpus data.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20K00583.

References

Marie-Catherine de Marneffe, Bill MacCartney and Christopher Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. LREC *2006*.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre and Dan Zeman. 2021. Universal Dependencies. Computational Linguistics, 47 (2): 255-308. https://doi.org/10.1162/COLI_a_00402.

Richard Futrell, Kyle Mahowald and Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. Proceedings of Natural Academy of Science, 112(33):10336-10341. https://doi.org/10.1073/pnas.1502134112

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95-126). MIT Press, Massachusetts, US.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? Cognitive Science, 34(2):286-310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science, 29(2):261-290. https://doi.org/10.1207/s15516709cog0000_7

Shin-Ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Learner corpus studies in Asia and the world, 1: 91-118.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. Language Sciences, 50: 93-104. https://doi.org/10.1016/j.langsci.2015.04.002Jiang, J. & H.

Jingyang Jiang and Haitao Liu (eds). 2018. Quantitative Analysis of Dependency Structures. Berlin: De Gruyter Mouton.

Jingyang Jiang and Jinghui Ouyang. 2017. Dependency distance: A new perspective on the syntactic development in second language acquisition. Physics of Life Review, 21: 209-210.

Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. Journal of Quantitative Linguistics 27(1): 62-79.

Wenping Li and Jianwei Yan. 2021. Probability distribution of dependency distance based on a Treebank of Japanese EFL learners' Interlanguage. Quantitative Linguistics, 28(2): 172-186, DOI: 10.1080/09296174.2020.1754611

Haitao Liu. 2007. Probability distribution of dependency distance. Glottometrics, 15:1-12.

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

Haitao Liu. 2009. Probability distribution of dependencies based on Chinese dependency treebank.

Journal of Quantitative Linguistics, 16(3): 256-273. https://doi.org/10.1080/09296170902975742

Haitao Liu, Chunshan Xu and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 21:171-193. https://doi.org/10.1016/j.plrev.2017.03.002

Jinghui Ouyang and Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? Journal of Quantitative Linguistics, 25(4): 295-313. https://doi.org/10.1080/09296174.2017.1373991

Masanori Oya. 2013. Degree centralities, closeness centralities, and dependency distances of different genres of texts. Selected papers from the 17th Conference of Pan-pacific Association of Applied Linguistics, 42-53.

Masanori Oya. 2021. Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus. Pacific Asia Conference on Language, Information and Computation (PACLIC) 35, Nov 7, 2021

Ioan-Iovitz Popescu, Karl-Heinz Best and Gabriel Altmann. 2014. Unified modeling of length in language. Studies in Quantitative Linguistics 16. RAM-Verlag.

Yalan Wang. 2020. Quantitative Analysis of Dependency Structures. Journal of Quantitative Linguistics 27(1): 83-91.

Dan Zeman. 2015. Slavic Languages in Universal Dependencies. Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning. Slovakia.

Dan Zeman and Joakim Nivre, et al. 2020. Universal Dependencies 2.7, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3424.