Cross-Linguistic Variances of Dependency Distances in Multi-Lingual Parallel Corpus

Masanori Oya

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Cross-Linguistic Variances of Dependency Distances in Multi-Lingual Parallel Corpus

Masanori Oya
School of Global Japanese Studies
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This study investigates whether there are differences in the variances of dependency distances of all dependency types across different languages, and also whether there are differences in the variances of dependency distances of core dependency types (nominal subjects, objects, and obliques) across different languages. Statistical tests using a multi-lingual parallel corpus data indicate that there are significant cross-linguistic differences in the variances of dependency distances of different dependency types, yet some language pairs do not show statistically significant differences.

1 Background

The theoretical background of this study is *Dependency Grammar* (*DG*) (Tesnière 1959). In the framework of DG, every word in a sentence depends on another word in the same sentence, and the main verb of the sentence depends on nothing. For example, in the sentence *David read 30 articles for his term paper*, the noun *David* depends on the verb *read* as the subject; the verb *read* depends on no other words in this sentence; the numeral *30* depends on the noun *articles*; the noun *articles* depends on the verb *read* as its object, etc. These dependencies are categorized into different *types*. For example, in the framework of *Universal Dependencies* (UD) (Zeman et al. 2017), the dependency between the noun *David* and the verb *read* is *nsubj* (nominal subject), between the verb *read* and the noun *article* is *obj* (direct object), etc.

Dependency distance (DD) is the number of words from a word in a sentence to the word which depends on the word. For example, in the example sentence above, the DD between the noun *David* and the verb *read* is one; the DD between the numeral *30* and the noun *article* is two; the DD between the noun *article* and the verb *read* is two.

DD attracts many researchers' attention as one of the measures for syntactic complexity (Gibson, 1998, 2000; Gildea and Temperley, 2010; Grodner and Gibson, 2005; Li and Yan, 2021; Liu, 2007, 2008; Liu et al., 2017). Some researchers have argued for the idea that DD represents a certain aspect of the universal properties of natural languages (Ouyang and Jiang, 2018; Ouyang, Jiang and Liu 2022; Wang and Liu 2017; Yang and Li 2019, among others). It is also argued that there is a cross-linguistic preference for shorter DDs due to the limit of short-term memory (*Dependency-Distance Minimization*) (Gibson 2000; Gildea and Temperley 2010; Temperley 2007, 2008, among others).

One of the most unique research programs related to DDs is curve-fitting of the frequency distributions of DDs. Previous studies have discovered that the frequency distribution of DDs can fit well with the right truncated modified Zipf-Alekseev Distribution (ZAD) (Jiang and Liu, 2015; Liu, 2009; Ouyang and Jiang, 2018). Frequency distributions of DDs across different languages also fit well with ZAD, and cross-linguistic variations are represented by different settings of the two parameters of ZAD (Niu, Wang and Liu 2023).

Even though we cannot deny the fact that fitting of the frequency distributions of DDs across languages provides us with a unique and promising field of investigation, it is also certain that there can be cross-linguistic differences among these frequency distributions of DDs which are also of linguistic value. Provided that different settings of the two parameters of ZAD

can indicate cross-linguistic differences of how well they fit with ZAD, it is difficult to interpret these parameters. For example, what does it mean when the parameter a of Language A is larger than that of Language B, and vice versa? In addition to this, it can be assumed that frequency distributions of dependency distances of different *dependency types* are different from each other, yet this assumption cannot be tested by curve-fitting of the frequency distribution of all the dependency distances of one language, and we may need to investigate the behaviors of dependencies of different dependency types in different languages, in order to understand them deeper than now.

2 This study

This study aims to statistically test whether there are differences in the variances of DDs for all dependency types across different languages, as well as for specific dependency types. These tests are expected to confirm that the variances of DDs exhibit significant cross-linguistic differences and that differences in dependency types will be reflected in the variances of DDs. The results of these tests will deepen our understanding of DDs. The research questions of this study are as follows:

1. Are variances of DDs of all the dependency types different across different languages?
2. Are variances of DDs of some dependency types different across different languages?

2.1 Data

The data used in this study come from the *Parallel Universal Dependencies Treebanks* 2.7 (PUD). The details of PUD are available at the Web page of the shared task on Multilingual Parsing from Raw Text to Universal Dependencies in CoNLL 2017 (http://universaldependencies.org/conll17/).

This study covers all the 21 languages in PUD: Arabic, Chinese, Czech, English, Finnish, French, Galician, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish. Each language in PUD has 1,000 sentences, translated from English sentences, and they have been annotated with morphological and syntactic tags which were provided by Google. They are further converted into *Universal Dependencies* (UD) (Zeman et al. 2017). The details of UD are available at its website (https://universaldependencies.org/).

The fact that the sentences in PUD are translation pairs across languages allows us to regard the syntactic differences (including differences in DDs) across them as being controlled in terms of their meanings.

2.2 Methodology

For each language in the PUD, 1,000 dependencies were randomly selected. Each of these dependencies has a unique DD. This random selection was repeated for all 21 languages in the PUD, resulting in 21 sets of 1,000 DDs (Set 1). Next, for each language in the PUD, the dependencies typed as *nsubj* (nominal subjects), *obj* (direct objects), and *obl* (noun phrases in the oblique case) were extracted, and these extractions were repeated for all 21 languages. This study focuses on these *core* dependency types because they represent the core arguments of predicates, and thus are expected to exhibit the central features of their behavior. Then, we have 21 sets of DDs typed as *nsubj* (Set 2), 21 sets of DDs typed as *obj* (Set 3), and 21 sets of DDs typed as *obl* (Set 4). The sizes of these sets vary across languages. The distribution of DDs is not expected to be normal, so non-parametric statistical tests should be conducted. Therefore, a Kruskal-Wallis test, one of such non-parametric tests, was conducted on each of the sets (Set 1, 2, 3, and 4) independently, using the web application *js-STAR XR+ release 2.1.2 j*, to examine whether their variances are statistically significantly different across the 21 languages.

2.3 Results

Tables 1, 2, 3, and 4 summarize the descriptive statistics of Sets 1, 2, 3, and 4, respectively. Across these sets, the majority of the means do not exceed four, and the medians and the modes are either one or two, indicating that these languages prefer short DDs, and across the three dependency types *nsubj*, *obj*, and *obl*, suggesting the effect of Dependency-Distance Minimization (Gibson 2000; Gildea and Temperley 2010; Temperley 2007, 2008, among others).

	Mean	Med.	Mod.	S.D.	Variance	N
ar	3.206	1	1	4.617	21.314	20439
cs	3.368	2	1	4.042	16.342	18406
de	4.143	2	1	4.883	23.844	21332
en	3.499	2	1	4.249	18.052	21030
es	3.404	2	1	4.574	20.922	23154
fi	3.154	2	1	3.494	12.209	15811
fr	3.478	2	1	4.755	22.613	24726
gl	3.403	2	1	4.660	21.718	23292
hi	4.235	2	1	5.571	31.032	23725
id	3.162	2	1	4.042	16.338	19345
is	3.196	2	1	3.957	15.659	17038
it	3.438	2	1	4.633	21.469	23569
ja	3.795	2	1	6.451	41.613	28788
ko	3.486	1	1	4.912	24.130	16488
pl	3.215	2	1	4.019	16.154	18384
pt	3.432	2	1	4.590	21.069	23277
ru	3.279	2	1	4.118	16.957	19355
sv	3.313	2	1	4.047	16.379	19052
th	2.699	1	1	3.283	10.781	22289
tr	3.537	1	1	4.660	21.718	16720
zh	4.003	2	1	5.004	25.037	21407

Table 1: Descriptive statistics of the DDs of all dependency types in 21 languages of PUD (Set 1); Med.; median: Mod.; mode: ar; Arabic: cs; Czech: de; German: en; English: fi; Finnish: fr; French: gl; Galician: hi; Hindi: id; Indonesian: is; Icelandic: it; Italian: ja; Japanese: ko; Korean: pl; Polish: pt; Portuguese: ru; Russian: sv; Swedish: th; Thai: tr; Turkish: zh; Chinese.

	Mean	Med.	Mod.	S.D.	Variance	N
ar	2.051	1	1	2.384	5.685	1511
cs	3.180	2	1	3.012	9.069	1398
de	4.539	3	1	4.230	17.893	1689
en	3.055	2	1	2.956	8.740	1631
es	3.728	2	1	3.724	13.864	1355
fi	2.336	2	1	2.018	4.071	1475
fr	3.737	2	1	3.887	15.105	1621
gl	3.618	2	1	3.735	13.953	1342
hi	7.719	6	2	6.035	36.417	1294
id	2.723	2	1	2.599	6.756	1936
is	2.592	2	1	2.606	6.791	1795
it	3.998	3	2	3.932	15.460	1293
ja	10.066	7	2	9.145	83.625	1517
ko	5.909	4	1	5.797	33.607	1706
pl	3.159	2	1	2.959	8.757	1175
pt	3.660	2	1	3.710	13.761	1490
ru	2.746	2	1	2.974	8.845	1548
sv	2.481	1	1	2.533	6.414	1765
th	3.292	2	1	3.247	10.546	1689
tr	7.128	5	1	5.976	35.708	1239
zh	4.260	2	1	4.440	19.716	1843

Table 2: Descriptive statistics of the DDs of the dependency type *nsubj* in 21 languages of PUD (Set 2); Med.; median: Mod.; mode: ar; Arabic: cs; Czech: de; German: en; English: fi; Finnish: fr; French: gl; Galician: hi; Hindi: id; Indonesian: is; Icelandic: it; Italian: ja; Japanese: ko; Korean: pl; Polish: pt; Portuguese: ru; Russian: sv; Swedish: th; Thai: tr; Turkish: zh; Chinese.

The Kruskal-Wallis test on Set 1 indicated a significant difference of means among them ($H = 183.04$, $p < .01$). Steel-Dwass tests were conducted to test pairwise comparisons between all the possible language pairs ($n = (21*21-21)/2 = 210$), and 56 pairs were found to be significantly different (about 26% of all the possible language pairs). This means that the majority of the language pairs show similar variances of DDs of all dependency types.

Among the Germanic languages in the PUD (English, German, Icelandic, and Swedish), significantly different pairs are German and Icelandic, and German and Swedish. All the possible pairs of Romance languages in PUD (French, Galician, Italian, Portuguese, and Spanish) are not significantly different. All the possible pairs of Slavic languages in PUD (Czech, Polish, and Russian) are also not significantly different. Variances of several language pairs are not significantly different even though they do not belong to the same language branch or are not used in geographically adjacent areas (e.g., Arabic and Japanese, German and Hindi, Icelandic and Indonesian, Italian and Korean).

Word-order patterns seem to be related to the results of the tests. The variances of Japanese and of Turkish (both are SOV languages) are significantly different from those of 19 other languages except for Arabic (a VSO language); Hindi and Korean (both SOV languages) are significantly different from all the other 20 languages.

The Kruskal-Wallis test on Set 2 showed that there was a significant difference of means among them ($H = 4633.55$, $p < .01$), and Steel-Dwass tests for all the possible language pairs show that 108 pairs (about 51% of all the possible language pairs) were significantly different. All 6 pairs of Germanic languages in PUD are significantly different, while all 10 pairs of Romance languages in PUD are not significantly different. As for Slavic languages in PUD, only the pair

	Mean	Med.	Mod.	S.D.	Variance	N
ar	1.957	1	1	1.938	3.757	746
cs	2.125	2	1	1.710	2.925	744
de	3.610	3	1	2.954	8.726	898
en	2.212	2	2	1.150	1.322	876
es	1.985	2	2	1.026	1.053	785
fi	2.001	2	1	1.459	2.129	924
fr	2.177	2	2	1.264	1.598	1082
gl	2.169	2	2	1.327	1.761	933
hi	2.626	1	1	3.103	9.628	1469
id	1.186	1	1	0.499	0.249	857
is	1.824	1	1	1.432	2.050	824
it	2.178	2	2	1.062	1.128	849
ja	2.807	2	2	2.585	6.683	843
ko	1.717	1	1	2.148	4.615	1030
pl	1.750	1	1	1.290	1.665	815
pt	2.102	2	2	1.155	1.334	882
ru	1.704	1	1	1.023	1.046	749
sv	2.352	2	1	1.796	3.225	900
th	1.254	1	1	1.363	1.858	1734
tr	2.205	1	1	2.684	7.203	1085
zh	3.407	3	1	2.830	8.008	1528

Table 3: Descriptive statistics of the DDs of the dependency type *obj* in 21 languages of PUD (Set 3); Med.; median: Mod.; mode: ar; Arabic: cs; Czech: de; German: en; English: fi; Finnish: fr; French: gl; Galician: hi; Hindi: id; Indonesian: is; Icelandic: it; Italian: ja; Japanese: ko; Korean: pl; Polish: pt; Portuguese: ru; Russian: sv; Swedish: th; Thai: tr; Turkish: zh; Chinese.

	Mean	Med.	Mod.	S.D.	Variance	N
ar	4.021	3	2	3.429	11.755	2133
cs	3.496	3	2	2.893	8.372	1348
de	4.207	3	1	3.671	13.479	1544
en	4.894	4	3	3.226	10.407	1275
es	4.462	3	3	3.400	11.561	1713
fi	2.967	2	1	2.238	5.006	1456
fr	5.140	4	3	3.948	15.586	1541
gl	4.926	4	3	3.772	14.229	1457
hi	7.528	6	2	5.919	35.038	2002
id	3.740	3	2	2.988	8.927	1398
is	3.816	3	2	2.577	6.643	1349
it	4.813	3	3	3.537	12.510	1617
ja	7.184	4	2	7.453	55.540	1647
ko	3.837	2	1	4.309	18.568	1973
pl	3.855	3	2	2.814	7.921	1550
pt	4.954	3	3	3.775	14.252	1424
ru	3.826	3	2	2.943	8.660	1477
sv	4.165	3	2	2.836	8.043	1326
th	3.794	3	2	3.042	9.256	1760
tr	4.369	2	1	4.811	23.150	1465
zh	4.339	2	1	4.929	24.295	961

Table 4: Descriptive statistics of the DDs of the dependency type *obl* in 21 languages of PUD (Set 4); Med.; median: Mod.; mode: ar; Arabic: cs; Czech: de; German: en; English: fi; Finnish: fr; French: gl; Galician: hi; Hindi: id; Indonesian: is; Icelandic: it; Italian: ja; Japanese: ko; Korean: pl; Polish: pt; Portuguese: ru; Russian: sv; Swedish: th; Thai: tr; Turkish: zh; Chinese.

Czech and Polish is not significantly different. Like the test results on Set 1, several language pairs are not significantly different even though they do not belong to the same language branch and they are used in geographically distant areas.

The Kruskal-Wallis test on Set 3 showed that there was a significant difference of means among them ($H = 3972.53$, $p < .01$). Steel-Dwass tests were conducted to test pairwise comparisons, and it was found that 158 pairs were significantly different (about 75% of all the possible language pairs). The four Romance languages in PUD are not significantly different among themselves; only the pair of French and Spanish is significantly different. The three Slavic languages in PUD are not significantly different among themselves. The four Germanic languages in PUD are significantly different among themselves; only the pair of English and Swedish is not significantly different.

The Kruskal-Wallis test on Set 4 showed that there was a significant difference in means among them ($H = 2675.99$, $p < .01$). Steel-Dwass tests were conducted to test pairwise comparisons, and it was found that 149 pairs were significantly different. Several language pairs are not significantly different even though they do not belong to the same language branch. Results are divided within Germanic languages in PUD: Of the possible six pairs, three of them are significantly different (German vs. English, English vs Icelandic, and English vs. Swedish), while three others are not (German vs. Islandic, German vs. Swedish, and Islandic vs. Swedish). Of the possible 10 pairs of Romance languages in PUD, only two pairs are significantly different (French vs. Spanish, Galician vs. Spanish).

3 Discussion

These results described above suggest that Romance languages seem to share similar properties in terms of the variances of dependency distances, yet the variances of

dependency distances of other languages do not suggest any correlation between which language branch they belong to and the variance of dependency distances.

The result that the majority of the language pairs show similar variances of DDs of all the dependency types does not seem to contradict the results of the previous studies on curve-fitting of the frequency distributions of DDs with ZAD. However, the tests of the variances of DDs of different dependency types show noteworthy differences across the languages in the corpus data. These results would not be captured appropriately only by curve fitting of frequency distributions of DDs of all dependency types with ZAD.

We may deepen our understanding of their distributions by testing the variances of the DDs of each of all the other dependency types, to ascertain which dependency types show more cross-linguistic variations than other dependency types.

In addition to this, we may curve-fit with ZAD the frequency distributions of the DDs of not only the core dependency types, but also each of other types, in the same corpus data, to ascertain how well they fit with ZAD. By doing this, we may have some insight into how we can interpret the settings of the parameters of ZAD across different languages and different dependency types, which will be one of the questions of future research.

4 Conclusion

This study attempted to test statistically whether there are differences in the variances of dependency distances of all dependency types across different languages, and also whether there are differences in the variances of dependency distances of core dependency types across languages. The statistical test results indicated that there are significant cross-linguistic differences in the variances of dependency distances in the languages in a multi-lingual parallel corpus. Further studies are required for a better understanding of cross-linguistic variation of dependency distances, while also focusing on their similarities.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP24K04089.

References

Yu Fang and Haitao Liu. 2018. What factors are associated with dependency distances to ensure easy comprehension? A case study of ba sentences in mandarin Chinese. *Language Sciences*, 67, 33–45. https://doi.org/10.1016/j.langsci.2018.04.005

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. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, A.P, Miyashita, W. 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

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.2015.04.00

Wenping Li and Jianwei Yan. 2021. Probability distribution of dependency distance based on a treebank of Japanese EFL learners' interlanguage. *Journal of Quantitative Linguistics*, 28(2), 172-186. https://doi.org/10.1080/09296174.2020.1754611

Haitao Liu. 2007. Probability distribution of dependency distance. *Glottometrics*, 15(1), 1-12.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159-191. https://doi.org/10.17791/jcs.2008.9.2.159

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 Junyin 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
Ruochen Niu, Yaqin Wang, and Haitao Liu. 2023. The cross-linguistic variations in Dependency Distance Minimization and its potential explanations. *Proceedings of 37th Pacific Asia Conference on Language, Information and Computing (PACLIC)*.
Jinghui Ouyang and Jingyang Jiang. 2018. 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
Jinghui Ouyang, Jingyang Jiang and Haitao Liu. 2022. Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing*, 51, 100-603. https://doi.org/10.1016/j.asw.2021.100603
David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2), 300–33. https://doi.org/10.1016/j.cognition.2006.09.011
David Temperley. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3), 256–82. https://doi.org/10.1080/09296170802159512
Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. Language Sciences 59, 135–147. https://doi.org/10.1016/j.langsci.2016.09.006
Hengbin Yan and Yanghui Li. 2019. Beyond length: Investigating dependency distance across L2 modalities and proficiency levels. *Open Linguistics*, 5(1), 601–614. https://doi.org/10.1515/opli-2019-0033
Daniel Zeman. 2015. Slavic languages in Universal Dependencies. *Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning*. 151-163.
Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,* pages 1–19, Vancouver.