Low-Frequency Long-Distance Dependencies as “Long Tails”

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

Abstract

This study focuses on long dependency distances in different languages, and proposes that they can be treated as a type of “long tail” for explicating minute yet significant language-specific variations in a numerical and objective way. It is found that Japanese prefers longer dependency distances, and different languages have different preferences for dependency types in long dependency distances. Such findings are expected to be applied to language-specific tuning-up for the output of language-generating applications.

1 Introduction

This study focuses on *long dependency distances* (henceforth LDD) in different languages, and proposes that they can be treated as a type of “long tail” for explicating minute yet significant language-specific variations in a numerical and objective way. Such findings are expected to be applied to language-specific tuning-up for the output of language-generating applications.

The long tail as we know it today refers to a type of business strategy first introduced by Chris Anderson in 2004 (published later in Anderson (2006)), which advocates selling a variety of items of low demand to many customers, thus gaining market share collectively which is larger than that gained by selling only items of high demand. The essence of the idea of long tail is that items or events with low frequency can have a significant influence on the phenomenon of interest.

The long tail in the context of *dependency distances* (henceforth DDs) is that we can explicate language-specific characteristic by focusing on DDs which are much longer than average. It is expected that these low-frequency DDs show language-specific characteristics more conspicuously than focusing on high-frequency ones.

The term LDD is not the same as long-distance dependencies that we find in generative syntax for wh-movement. The definition of LDD in this study does not necessarily involve wh-movement. For example, in the sentence (1) below (chosen from *English Parallel Universal Dependency* Universal Dependency (Zeman et al. 2017), sentence id; w01075037), the verb of the subordinate clause (*resulted*) depends on the verb of the main clause (*reduced*), and the DD is long because there are 21 words between them. On the other hand, the subject noun (*enrollment*) depends on the verb of the main clause, and the DD is short because there are five words between them (the number “1”, “3”, “10”, commas, and the symbol “%” are counted as individual words):

(1) Specifically, a male secondary school enrollment 10% above the average *reduced* the chance of a conflict by about 3%, while a growth rate 1% higher than the study average *resulted* in a decline in the chance of a civil war of about 1%.

2 Background

It has already been revealed that there is a cross-linguistic tendency that DDs which are four or less are more frequent than those which are five or more, and that the average DD is less than four across different languages (Fang and Liu 2018; Futrell et al. 2015; Gibson, 1998, 2000; Gildea and Temperley, 2010; Grodner and Gibson, 2005; Li and Yang 2021; Liu 2007, 2008; Liu et al., 2017; Ouyang and Jiang, 2018; Ouyang, Jiang and Liu 2022; Wang and Liu 2017; Yang and Li 2019). The cause of this tendency has been argued to be the result of the capacity of the working memory of humans when producing a sentence (Gibson 2000; Gildea and Temperly 2010; Temperly 2007, 2008, among others); longer DDs yield heavier burdens on memory than shorter ones, hence shorter DDs are preferred to longer ones in order to ensure efficient processing of sentence generation.

Along with the preference of natural languages for shorter DDs and focus on these frequent cases, it is worthy to focus on LDDs, which are rare cases, viz a long tail. When we plot the DDs of two languages on x axis and their frequencies on y axis, then we have their frequency distributions shown in a model distribution below:

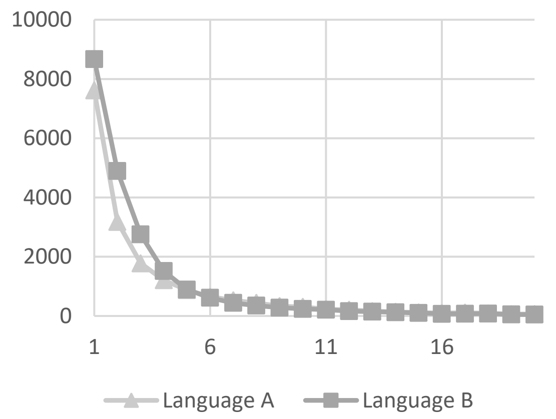

Figure 1: Model frequency distributions of dependency distances of two languages

When the focus of investigation is the shorter DDs of higher frequency, it is natural to ignore DDs much longer than average, and this stance has been chosen in many studies in DDs so far. When, on the other hand, the focus of investigation is explicating the characteristics of different languages, or their language-specific naturalness represented by a variety of variables (DD is assumed here to be one of them), we cannot ignore any variable only because their frequencies are low. Rather, rare cases (such as the long tails illustrated above) can show us something valuable in the context of characterizing given languages through a variety of variables, because these rare cases in sum can contribute to characterize the language in a unique way.

In addition to this, in terms of information theory (Shannon 1948), the content of information is larger in rare cases than in common, more frequent cases. This means that LDDs are expected to convey more information than short DDs, therefore focusing on LDDs means focusing on more informative aspect of human languages. It is safe to say that this viewpoint has not necessarily been stressed in the trend of NLP research, and therefore it is worth focusing on.

One example of such investigation focusing on rare cases in natural languages is an investigation of one-word sentences in the Japanese sentences and how they are translated into English sentences in English-Japanese parallel corpus (Oya 2015).

In the context of focusing on LDDs, what seems to be interesting is the dependency types which are used in LDDs and which are not. If it is found that a certain dependency type in one language is used in LDDs more frequently than other types in the same language, or across different languages, these results can contribute to explicate the characteristics of these languages in focus in a unique way. For example, it is expected that the dependency type *nsubj* (the dependency type of the dependency between a verb and its subject; all the dependency type names used in this study are based on Universal Dependency (Zeman 2015; Zeman et al. 2017)) is frequently used in short dependency distances in English but not necessarily in Japanese, because the dependency distance between the subject and the verb in SVO languages must be shorter than that in SOV languages. On the other hand, it is expected that the frequencies of short dependency distance of the type *obj* (the dependency type between a verb and its direct object) are similar across English and Japanese, and those of long dependency distance of the same type are low across these languages.

3 This study

The background described above has motivated me to investigate the long tails of the frequencies of LDDs, and the research questions of the current research are as follows: (1) Do different languages show different tendencies in terms of the frequencies of LDDs? (2) Are specific dependency types used more frequently in LDDs?

3.1 Data

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

Among the 20 languages in PUD, this study focuses on the following four languages: Mandarin Chinese, English, Japanese, and Korean. Each language in PUD has 1,000 parallel sentences, translated from the original English sentences. These sentences were annotated with morphological and syntactic tags by Google, which are further converted into Universal Dependencies (henceforth UD) by UD community members, according to UD Ver. 2 guidelines. The UD website elucidates further details (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.

3.2 Methods

The dependency distances of all the dependency relationships in these four languages in PUD are counted using a filtering function of a spreadsheet application, and their frequencies are plotted using the same application. The format of PUD allows us to use a simple filtering function of a spreadsheet application to count the number of DDs within the following domains:

Domain 1: DD 10 or shorter
Domain 2: DD longer than 10, and 20 or shorter
Domain 3: DD longer than 20, and 30 or shorter
Domain 4: DD longer than 30, and 40 or shorter

Then, the distributions of DDs within each of the four domains above are compared across these four languages, using Kruskal-Wallis tests (the Web application used for these tests is *js-STAR XR+ release 1.6.0 j*), so that we can ascertain if a particular language has (or languages have) different distribution(s) of longer dependency distances; LDDs longer than 40 are ignored in this study because their frequencies are too small to draw any conclusion from the results:

In addition to this, using a filtering function of the spreadsheet application, we count the 10 most frequently used dependency types in these four domains of these four languages, so that we can find out which dependency types are more frequently used in LDDs in which language.

3.3 Results

The figure below is the frequency distributions of dependency distances 40 or shorter (all the 4 domains):

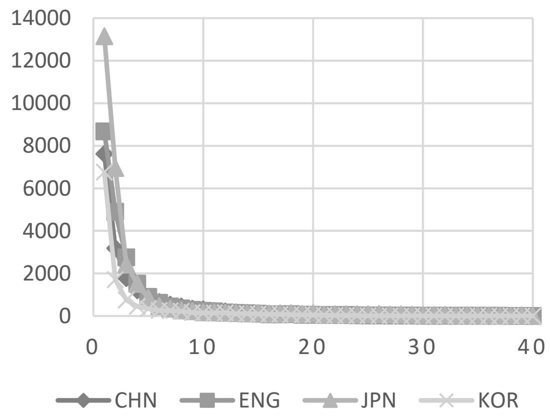

Figure 2: The frequency distributions of the dependency distances 40 or shorter of the four languages in PUD;

The distributions of DDs are quite similar across these four languages, and the majority of DDs fall below 4, as expected from the previous studies of DDs.

The followings figures are the frequency distributions of dependency distances of these 4 domains:

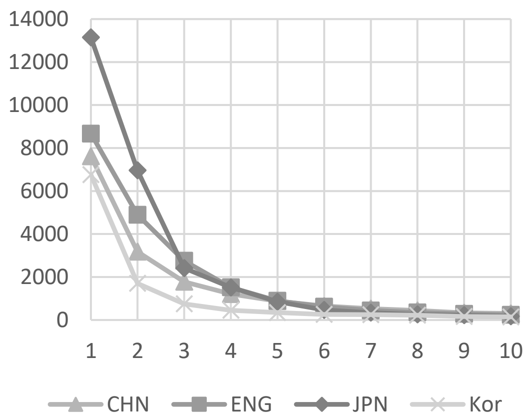

Figure 3: The frequency distributions of the dependency distances of the four languages in PUD in the Domain 1 (DD 10 or shorter)

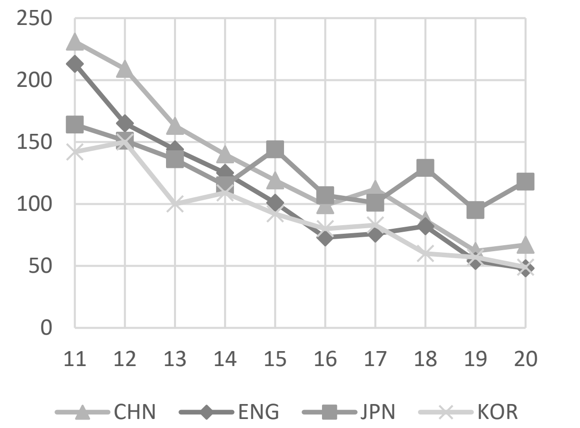

Figure 4: The frequency distributions of the dependency distances of the four languages in the Domain 2 (D.D. longer than 10, and 20 or shorter)

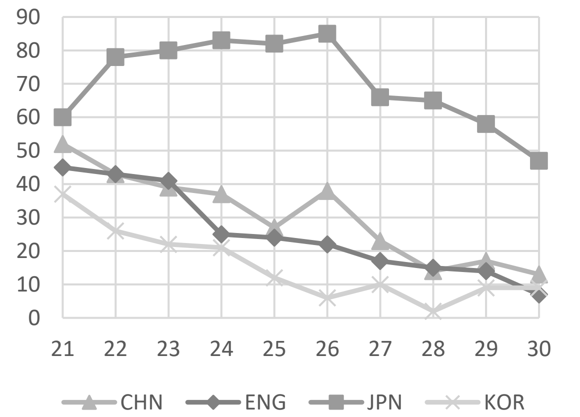

Figure 5: The frequency distributions of the dependency distances of the four languages in the Domain 3 (D.D. longer than 20, and 30 or shorter)

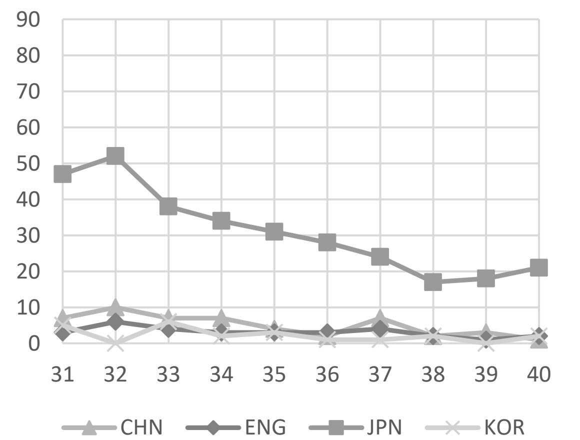

Figure 6: The frequency distributions of the dependency distances of the four languages in the Domain 4 (D.D. longer than 30, and 40 or shorter)

The ranges of the y axes are different across these domains, yet as far as Domain 3 and 4 are concerned, the LDDs of Japanese show distributions which are different from those of the other 3 languages. This characteristic would be ignored easily if we only observe the whole distributions of DDs, like that shown in Figure 2.

The results of Kruskal-Wallis tests on these data show that, both in Domain 1 and Domain 2, there are no statistically significant differences among the frequencies of DDs across these four languages. In Domain 3 and 4, as expected from the distribution graphs above, there are statistically significant differences among them. Therefore, for these domains, multiple comparisons are conducted using Steel-Dwass tests, and it is found out that the frequencies of DDs of Japanese are different from all the other three languages in PUD statistically significantly.

Table 1 shows us the 10 most frequently used dependency types of these four languages. These frequently-used dependency types are quite similar across these four languages, though their ranks vary from language to language.

Chinese (n= 21407)			English (n=21026)		
Types	Freq.	Ratio	Types	Freq.	Ratio
punct	2894	0.135	case	2499	0.119
compound	1777	0.083	punct	2301	0.109
nsubj	1776	0.083	det	2047	0.097
obj	1526	0.071	nsubj	1393	0.066
advmod	1332	0.062	amod	1336	0.064
case	1319	0.062	obl	1237	0.059
root	1000	0.047	nmod	1076	0.051
nummod	809	0.038	root	1000	0.048
nmod	702	0.033	obj	876	0.042
aux	686	0.032	advmod	852	0.041

Japanese (n=22910)			Korean(n=16584)		
Types	Freq.	Ratio	Types	Freq.	Ratio
case	6489	0.283	compound	2282	0.138
punct	3028	0.132	obl	1869	0.113
aux	2701	0.118	punct	1595	0.096
nmod	2092	0.091	nsubj	1546	0.093
obl	1596	0.070	acl:relcl	1188	0.072
nsubj	1455	0.064	obj	1030	0.062
acl	1090	0.048	root	1000	0.060
root	1000	0.044	advcl	999	0.060
advcl	917	0.040	nmod:poss	655	0.039
obj	844	0.037	advmod	593	0.036

Table 1: The 10 most frequently used dependency types of Chinese, English, Japanese and Korean, their frequencies, and their ratios to all the dependency types of each language

The tables 2 through 4 (Domain 3 and 4 are merged, due to the scarcity of the data) show us the 10 most frequent dependency types of these four languages, their frequencies, and their ratios to all the dependency types across these four domains. Dependency types in bold are those not in the 10 most frequent dependency types in each language, and the number in parentheses is its rank in all the dependency types in each language.

Chinese (n= 21407)			English (n=21026)		
Types	Freq.	Ratio	Types	Freq.	Ratio
punct (1)	2547	0.119	case (1)	2498	0.119
compound (2)	1774	0.083	det (3)	2047	0.097
nsubj (3)	1615	0.075	punct (2)	1581	0.075
obj (4)	1487	0.069	nsubj (4)	1350	0.064
case (6)	1306	0.061	amod (5)	1336	0.064
advmod (5)	1252	0.058	obl (6)	1157	0.055
nummod (8)	809	0.038	nmod (7)	1068	0.051
nmod (9)	688	0.032	obj (9)	876	0.042
aux (10)	676	0.032	advmod (10)	837	0.040
mark:rel (11)	623	0.029	**compound (11)**	810	0.039

Japanese (n=22910)			Korean(n=16584)		
Types	Freq.	Ratio	Types	Freq.	Ratio
case (1)	6486	0.283	compound (1)	2279	0.137
punct (2)	3027	0.132	obl (2)	1731	0.104
aux (3)	2700	0.118	punct (3)	1571	0.095
nmod (4)	2080	0.091	nsubj (4)	1237	0.075
obl (5)	1349	0.059	acl:relcl (5)	1183	0.071
acl (7)	1068	0.047	obj (6)	1019	0.061
nsubj (6)	1038	0.045	advcl (8)	851	0.051
obj (10)	831	0.036	nmod:poss (9)	651	0.039
advcl (9)	658	0.029	advmod (10)	520	0.031
mark (11)	383	0.017	**nummod (11)**	487	0.029

Table 2: The 10 most frequently used dependency types of Chinese, English, Japanese and Korean in Domain 1, their frequencies, and their ratios to all the dependency types of each language

Chinese (n= 21407)			English (n=21026)		
Types	Freq.	Ratio	Types	Freq.	Ratio
root (7)	355	0.017	punct (2)	469	0.022
punct (1)	274	0.013	root (8)	180	0.009
nsubj (3)	141	0.007	**conj (12)**	87	0.004
dep (18)	138	0.006	obl (6)	75	0.004
advcl (13)	94	0.004	**advcl (18)**	63	0.003
advmod (5)	66	0.003	nsubj (4)	41	0.002
xcomp (14)	58	0.003	**parataxis (29)**	40	0.002
ccomp (17)	55	0.003	advmod (10)	14	0.001
obl (12)	39	0.002	**mark (14)**	11	0.001
obj (4)	37	0.002	**acl:relcl (24)**	11	0.001

Japanese (n=22910)			Korean(n=16584)		
Types	Freq.	Ratio	Types	Freq.	Ratio
root (8)	375	0.016	root (7)	508	0.031
nsubj (6)	277	0.012	nsubj (4)	253	0.015
advcl (9)	205	0.009	advcl (8)	125	0.008
obl (5)	182	0.008	obl (2)	121	0.007
cc (15)	40	0.002	advmod (10)	56	0.003
advmod (12)	27	0.001	**conj (14)**	27	0.002
nsubj: outer (17)	27	0.001	**obl: tmod (21)**	23	0.001
compound (14)	20	0.001	punct (3)	21	0.001
acl (7)	18	0.001	**nsubj: pass (20)**	9	0.001
nmod (4)	11	0.000	obj (6)	7	0.000

Table 3: The 10 most frequently used dependency types of Chinese, English, Japanese and Korean in Domain 2, their frequencies, and their ratios to all the dependency types of each language

Chinese (n= 21407)			English (n=21026)		
Types	Freq.	Ratio	Types	Freq.	Ratio
root (7)	355	0.017	punct (2)	469	0.022
punct (1)	274	0.013	root (8)	180	0.009
nsubj (3)	141	0.007	**conj (12)**	87	0.004
dep (18)	138	0.006	obl (6)	75	0.004
advcl (13)	94	0.004	**advcl (18)**	63	0.003
advmod (5)	66	0.003	nsubj (4)	41	0.002
xcomp (14)	58	0.003	**parataxis (29)**	40	0.002
ccomp (17)	55	0.003	advmod (10)	14	0.001
obl (12)	39	0.002	**mark (14)**	11	0.001
obj (4)	37	0.002	**acl:relcl (24)**	11	0.001

Japanese (n=22910)			Korean(n=16584)		
Types	Freq.	Ratio	Types	Freq.	Ratio
root (8)	375	0.016	root (7)	508	0.031
nsubj (6)	277	0.012	nsubj (4)	253	0.015
advcl (9)	205	0.009	advcl (8)	125	0.008
obl (5)	182	0.008	obl (2)	121	0.007
cc (15)	40	0.002	advmod (10)	56	0.003
advmod (12)	27	0.001	**conj (14)**	27	0.002
nsubj:outer (17)	27	0.001	**obl:tmod (21)**	23	0.001
compound (14)	20	0.001	punct (3)	21	0.001
acl (7)	18	0.001	**nsubj:pass (20)**	9	0.001
nmod (4)	11	0.000	obj (6)	7	0.000

Table 4: The 10 most frequently used dependency types of Chinese, English, Japanese and Korean in Domain 3 and 4, their frequencies, and their ratios to all the dependency types of each language

These results show that lower-ranked, less frequently used dependency types are used more frequently in LDDs, and these types show specific differences across languages: For example, the dependency type parataxis is used more frequently in LDDs in English than other three languages; *dep* (unspecified dependency types) and *xcomp* (external-subject complement) are more frequent in Chinese LDDs; *nsubj:outer* (a nominal subject of a copular clause whose predicate is itself a clause) and *compound* are more frequent in Japanese LDD; and *obl:tmod* (oblique-case nominal phrases expressing temporal modification) is more frequent in Korean LDDs.

4 Discussion

These results lead us to the following suggestions. First, focusing on the "long tails," or on the less-frequent LDDs, reveals language-specific characteristics which we may overlook if we only focus on shorter DDs. If we focused only on shorter, more frequent DDs, we would find that these languages are similar in terms of how they use their dependency types, as shown in Table 1 above, and we might even use these facts to support for a certain aspect of linguistic universal. However, that focus on shorter DDs would lead us to ignore the fact that Japanese language uses LDDs more frequently than the other three languages, and the fact that different languages use different dependency types in LDDs. This study has shed a unique light on this aspect of DD investigations.

Some may argue that these results are something which have already been expected; the logic behind this must be like "different languages show different characteristics, and different dependency types are nothing other than examples such characteristics." This might be true, yet it does not explain (nor does this study, actually) the reason why the dependency types of short DDs shown in Table 1 are similar across different languages; why do these languages show a certain level of similarity in short DDs, while they show differences in LDDs? This is the topic of future study. In this context, what is also to be investigated is the reason why certain dependency types are frequently used in LDDs of one language but not in those of other languages.

Second, the focus on less-frequent LDDs of these four languages can easily be extended to other languages, e.g., the other languages in PUD with the same method. This multi-lingual extension of the focus on LDDs will provide us with a unique perspective on less-frequent, yet not ignorable phenomena of natural language.

Lastly, the understanding of multi-lingual LDDs can be integrated into large-scale language models for tuning-up of generative AIs; there seems to be a consensus at present among the general public that generative AIs can generate natural sentences at a surprising speed; yet the quality of the sentences they generate can further be improved from various viewpoints, and occasional and language-specific use of LDDs can be integrated to the output of generative AIs, so that their output will sound more natural and humanlike than now. This idea has not yet been substantiated and therefore we need to conduct further research to investigate how distributions of LDDs can be integrated into the output of generative AI applications, and whether people actually perceive sentences with occasional LDDs more natural than those without. These issues are to be investigated in future research.

5 Conclusion

This study focused on long dependency distances in different languages, and proposed that they can be treated as a type of "long tail" for explicating minute yet significant language-specific variations in a numerical and objective way. It is found that Japanese prefers longer dependency distances, and different languages have different preferences for dependency types in long dependency distances. Such findings are argued to be applied to language-specific tuning-up for the output of language-generating applications.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20K00583.

References

Chris Anderson. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, New York City, NY.

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

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, 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

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

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *Bell Systems Technical Journal*, Vol. 27, pp 379–423, 623-656.

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

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.