Comparing Dependency Distances of Esperanto and Other Languages in a Multi-Lingual Parallel Corpus

Masanori Oya
Meiji University
Tokyo
masanori_oya2019@meiji.ac.jp

Abstract

This study attempts to integrate Esperanto into the research of dependency distance, based on a small-scale multi-lingual parallel corpus of Manifesto de Prago, annotated with UD relations, to contribute to the development of research on Esperanto as a natural language. The mean dependency distance and the distribution of dependency distances of Esperanto are not significantly different from the majority of the 21 languages in the corpus, thus partially supporting the claim that Esperanto is no less natural than other natural languages.

Keywords: Esperanto, Dependency Distance, Multi-Lingual Parallel Corpus

1. Introduction

This study attempts to integrate Esperanto into the research of *Dependency Distance*, based on a small-scale multi-lingual parallel corpus, which has been constructed for this unique purpose.

Despite the origin of Esperanto as a constructed language (Zamenhof, 1887), several researchers today investigated its linguistics characteristics as one of the natural languages (Gledhill, 2000; Goodall, 2023; Liu, 2011; Koutny, 2015, among others) which can be the topic of theoretical linguistics, because it shares some linguistic characteristics with other natural languages. In this context, Oya (2025) has introduced a small-scale *Universal Dependencies* (UD) treebank based on the Esperanto version of *Manifesto de Prago* (Fettes, 1996) (MdP), an official document published online, advocating the status of Esperanto as an international auxiliary language. That treebank is intended to contribute to the development of research on Esperanto as a natural language.

If Esperanto is to be considered as one of the natural languages as argued by some researchers mentioned above, then we need to demonstrate that (1) it does not show any extraordinary trait which cannot be found in other natural languages, and that (2) it shares some common characteristics with other natural languages. In order to do so, we need to employ some metrics to compare and contrast Esperanto to other natural languages so that such investigations can be as objective and replicable as possible. In this context, this study focuses on the Dependency Distance of Esperanto and other languages as such a metric, based on the text data provided in MdP as a multi-lingual parallel corpus.

This article is organized as follows: Section 2 summarizes the concept of Dependency Distance, with a focus on the idea of *Mean Dependency Distance* (Liu, 2007, 2008, etc.), followed by the rationale of this study. Section 3 describes how the raw text data available online were parsed in the format of UD for this study, and how these data are analyzed in terms of their Dependency Distances. Section 4 reports the results of the analysis, which are further discussed from several viewpoints in Section 5, and Section 6 concludes this article.

2. Dependency Distance

Dependency Distance (DD) is the number of words from a word in a sentence to the word on which it depends (Hudson, 1995; Liu, 2007, 2008, among others). Liu (2008) states the definition of DD as follows:

> "let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_a and W_b, if W_a is a governor and W_b is its dependent, then the *dependency distance* (DD) between them can be defined as the difference $a - b$; ... in measuring dependency distance the relevant measure is the absolute value of dependency distance."

For example, in a sentence *Sarah read 30 books for her Ph.D dissertation*, the noun *Sarah* depends on the verb *read*, and its DD is one; the numeral *30* depends on the noun *book*, and its DD is one; the noun *books* depends on the verb *read*, and its DD is two, etc.

DD has attracted considerable scholarly attention as a key metric of syntactic complexity. A number of studies further contend that DD captures certain universal properties of natural languages (Ouyang and Jiang, 2018; Ouyang et al., 2022;

Wang and Liu, 2017; Yan and Li, 2019, among others). Moreover, cross-linguistic research has proposed a general preference for shorter dependency distances, attributed to constraints on short-term memory. This is an effect commonly described as Dependency Distance Minimization (Gibson, 2000; Gildea and Temperley, 2010; Jiang and Ouyang, 2018; Niu et al., 2023; Temberley, 2007; Temperley, 2008, among others).

In this context of research on dependency distance and its minimization, mean dependency distance (MDD) is defined as follows (Liu, 2008), where n means the word count of a sentence and DD_i means the dependency distance between the i-th word and its head:

$$\mathrm{MDD}(sentence) = \frac{1}{n-1}\sum_{i=1}^{n}|DD_i| \qquad (1)$$

Liu (2008) assumed "that the greater the MDD of a sentence, the more difficult the sentence" and "the greater the MDD of a text, the more difficult the text (or language)", and conducted a corpus-based research of the MDDs of 20 languages (Esperanto is not included), and revealed that all MDDs of these languages fall below 3.662.

Oya (2024) also conducted a research on multilingual parallel corpus to calculate the DDs of a variety of languages (Esperanto is not included) to examine whether the distributions of DDs are different across different languages. The results show that the majority of the language pairs show similar distributions of DDs, though their MDDs are diverse within a certain threshold. This suggests that, while MDDs may reflect language-dependent characteristics across different languages within a certain threshold, the distributions of DDs are stable across them.

This study is an extension of Oya (2024), employing MdP mentioned above as a multi-lingual parallel corpus. MdP is actually provided not only in Esperanto, but in a variety of languages, and therefore it can serve as a multi-lingual parallel corpus when annotated in a principled way, such as UD. At present, the Esperanto version of MdP has been annotated with UD (Oya, 2025) as mentioned above. This study introduces UD annotation on MdP in 22 languages, including Esperanto, and examines whether the MDD and the distribution of DDs in Esperanto are similar to those of these 21 languages. If it is found the case, then that will partially support the claim that Esperanto share a certain trait with other natural languages as far as the distribution of DDs is concerned.

This study is also an extension of Liu (2011), which conducted analyses of lexical and syntactic features of an Esperanto book (Zamenhof, 1907). Along with other lexical and syntactic properties which strongly suggest that Esperanto is a structurally natural language, it revealed that the MDD of the book is 3.85, which is longer than MDDs of other languages mentioned in Liu (2008). If we follow the logic of Liu (2008), that is, the greater the MDD of a text, the more difficult the text (or language), this result could lead us to conclude that Esperanto is more difficult than other languages, and it would be in contrast to the claim that Esperanto is no less natural than other languages, as far as relative difficulty of a language (represented by a longer MDD) is considerd to represent its relative lack of naturalness. However, the comparison of MDDs across Liu (2008) and Liu (2011) is not sound because they analyze different corpus data, and therefore the contents of the text are not controlled for syntactic analyses across them.

This insight on the necessity of controling the content of the text data motivates to investigate the MDD of Esperanto along with those of other languages using a multi-lingual paralell corpus as the primary data, in which the texts share the same content across different languages, and therefore we can focus on the difference or similarity of their syntactic properties.

Against the background described above, this study attempts to answer the following question: Are the MDD and the distribution of DDs in Esperanto similar to those of the other languages in a multi-lingual parallel corpus?

3. This study

3.1. Data

The corpus data used in this study is Esperanto MdP with UD annotations, and its multi-lingual versions available online (https://lingvo.org/prago). It is a manifesto which describes seven principles of Esperanto movement to advocate it as an international auxiliary language, published in the 81st World Esperanto Congress in Prague, Czech Republic in 1996. It is translated into 50 languages.

For example, the sentence (1) below is quoted from MdP followed by its English translation, its UD tree is Figure 1, and its UD analysis is shown in Table 1:

(1) *Ni asertas, ke lingva malegaleco sekvigas komunikan malegalecon je ĉiuj niveloj, inkluzive de la internacia nivelo.*
(We maintain that language inequality gives rise to communicative inequality at all levels, including the international level.)

3.2. Analysis

The Esperanto version of MdP has been manually annotated with UD tags in the format of

sent_id = prago-009
text = Ni asertas ke lingva malegaleco sekvigas komunikan malegalecon je ĉiuj niveloj, inkluzive de la internacia nivelo.
text_en = We maintain that language inequality gives rise to communicative inequality at all levels, including the international level.

1	Ni	ni	PRON	_	Case=Nom|Number=Plur|Person=1|PronType=Prs	2	nsubj
2	asertas	aserti	VERB	_	Mood=Ind|Tense=Pres|VerbForm=Fin	0	root
3	,	,	PUNCT	_	_	7	punct
4	ke	ke	SCONJ	_	_	7	mark
5	lingva	lingva	ADJ	_	Degree=Pos	6	amod
6	malegaleco	malegaleco	NOUN	_	Case=Nom|Number=Sing	7	nsubj
7	sekvigas	sekvigi	VERB	_	Mood=Ind|Tense=Pres|VerbForm=Fin	2	ccomp
8	komunikan	komunika	ADJ	_	Case=Acc|Degree=Pos|Number=Sing	9	amod
9	malegalecon	malegaleco	NOUN	_	Case=Acc|Number=Sing	7	obj
10	je	je	ADP	_	_	12	case
11	ĉiujn	ĉiu	DET	_	Case=Acc|Number=Plur|PronType=Tot	12	det
12	nivelojn	nivelo	NOUN	_	Case=Acc|Number=Plur	7	obl
13	,	,	PUNCT	_	_	14	punct
14	inkluzive	inkluzive	ADV	_	_	7	advmod
15	de	de	ADP	_	_	18	case
16	la	la	DET	_	Definite=Def|PronType=Art	18	det
17	internacia	internacia	ADJ	_	Case=Nom|Degree=Pos|Number=Sing	18	amod
18	nivelo	nivelo	NOUN	_	Case=Nom|Number=Sing	14	nmod
19	.	.	PUNCT	_	_	2	punct

Table 1: The UD annotations of the sentence (1); the fields *DEPS* and *MISC* are excluded

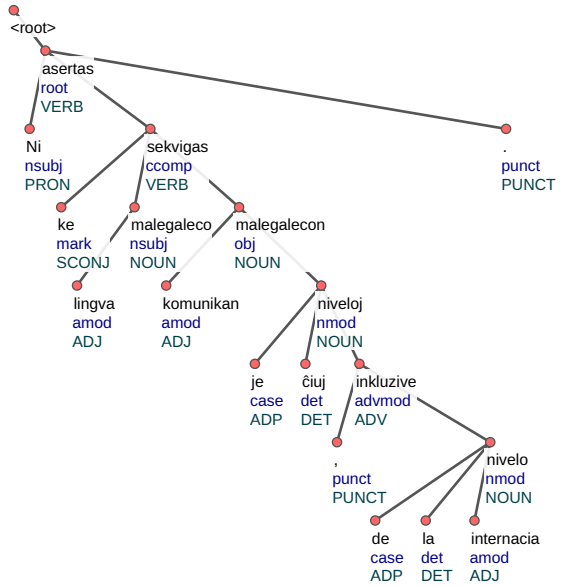

Figure 1: The UD tree for the sentence (1)

CoNNL-U and made public at the Web page of Universal Dependencies (Oya, 2025). Among these 50 languages in MdP, this study has chosen 20 languages, and parsed them using an original Python script which uses *SpaCy* for dependency parsing (I parsed the Hungarian version using *HuSpaCy*). The parsed output of these 20 languages has been made public at Github (https://github.com/MasanoriOya/Manifesto-de-Prago).

The DDs of the dependency relations in each language are extracted from the treebank, as the absolute value of the difference between the governor and its dependent, and its MDD is calculated using the formula proposed by Liu (2008). This process is repeated for all the 21 languages in the parallel corpus. Then these MDDs are compared and contrasted with eath other. The differences of the distribution of these DDs are statistically examined using a Kruskal-Wallis test. This test is chosen because the distribution of the DDs of a language is not normal. This means that we cannot use an ANOVA, which presupposes the normal distributions of the data.

4. Results

The descriptive statistics of the DDs of the 21 languages in MdP is shown in Table 2:

Multi-lingual comparisons of MDDs show us variances across these languages, which do not necessarily reflect their typological categorizations. In other words, languages which belong to the same branch are not necessarily close to each other in terms of their MDDs. The MDD of Esperanto is 3.409, which falls in the middle of the average dependency distances of these 21 languages, and it is also shorter than the result of Liu (2011) (3.85).

A Kruskal-Wallis test showed that the DDs significantly differed across these 21 languages, ($H(20) = 185.27, p < .01$). A post-hoc Steel-Dwass test revealed that the distribution of the DDs of Esperanto was significantly different from those of Catalan, English, Japanese and Spanish, but no significant difference was found between the distribution of DDs of Esperanto and those of other 16 languages.

5. Discussions

The result suggests that Esperanto is not conspicuously different from the majority of the languages in the MdP, as far as the MDD and the distribution of DDs is concerned. This fact may constitute a partial yet significant support of the claim that Esperanto is no less natural than other natural languages. This insight is available only from the viewpoint of Universal Dependencies and the DDs

	eo	ca	da	de	el	en
mean	3.41	3.29	2.99	3.72	3.08	2.85
median	2	2	2	1	2	1
mode	1	1	1	1	1	1
SD	4.60	4.99	4.31	5.39	4.57	4.36
token	839	1006	853	890	924	907

	es	fi	fr	hr	hu	it
mean	3.10	2.98	3.42	3.41	3.81	3.67
median	2	1	2	2	2	2
mode	1	1	1	1	1	1
SD	4.88	4.25	4.94	4.47	5.32	6.05
token	973	686	1055	785	913	1025

	ja	lt	nl	pl	pt	ro
mean	3.92	3.37	4.14	3.26	3.52	3.56
median	2	2	2	2	2	2
mode	1	1	1	1	1	1
SD	8.47	4.43	5.59	4.54	5.48	6.70
token	1432	737	926	794	963	864

	ru	sl	sv
mean	3.46	3.23	3.65
median	2	2	2
mode	1	1	1
SD	4.61	4.37	5.34
token	786	781	899

Table 2: The descriptive statistics of the dependency distances of the 21 languages: eo, Esperanto; ca, Catalan; da, Danish; de, German; el, Greek; en, English; es, Spanish; fi, Finnish; fr, French; hr, Croatian; hu, Hungarian; it, Italian; ja, Japanese; lt, Lithuanian; nl, Dutch; pl, Polish; pt, Portuguese; ro, Romanian; ru, Russian; sl, Slovenian; sv, Swedish

of different languages. It will be interesting to investigate whether Esperanto was actually less natural than other natural languages in the beginning yet it has come to have a characteristics which is similar to natural languages at present. For example, the longer MDD of Esperanto in the result of Liu (2011) may reflect the fact that its text data was Zamenhof (1907), which was published more than 100 years ago, while the result of this study is based on the parallel texts of MdP published in 1990s. This issue can be addressed by investigating the chronological change of Esperanto, based on larger-scale treebanks of Esperanto, which are not yet available today. Therefore, we need to construct them for that purpose along with others. This is one of the questions to be answered in future research.

Another issue of interest related to Esperanto is the difference between Esperanto and other constructed languages, such as Ido or Interlingua. If it is found that these are less natural than Esperanto in terms of the distribution of DDs, then it may serve as a partial support that Esperanto has come to be used more widely than these other constructed languages due to its naturalness. If, on the other hand, it is found that they are as natural as other languages like Esperanto, then it may support the claim that DDs represent a certain universal characteristics of language as we know it, whether it is natural or constructed. Again, we need large-scale corpora of these constructed languages for comparing and contrasting them to Esperanto, which is another question to be addressed in future.

Along with these issues related to Esperanto, this study collaterally found that some languages in MdP show significant differences from others. For example, the distribution of DDs of English is significantly different from those of Croatian, Dutch, Finnish, Hungarian, Italian, and Swedish, and the distribution of DDs of Dutch is significantly different from those of Catalan, English, Japanese, Polish, Portuguese, and Romanian. We need to explain why they are different from others, and that explanation may be applied to Esperanto, thus contributing to the understanding of why Esperanto is different from some natural languages as mentioned above.

Overall, this study only focuses on the distributions of DDs of Esperanto and other natural languages in a small-scale multi-lingual parallel corpus, and we need to address its limitations, such that (1) we need to employ other metrics for sentence complexity, and (2) we need larger-scale multi-lingual parallel corpus. As for (1), there are several metrics for it, such as the average sentence lengths of languages or the centrality of dependency structures. As for (2), *Parallel Universal Dependencies* (PUD) would be a good candidate; by constructing an Esperanto part of PUD and employing it for further research, we will deepen our understanding of the characteristics of this language which can be both unique and universal, just like other natural languages.

6. Conclusion

This study attempted to integrate Esperanto into the research of dependency distance, based on a small-scale multi-lingual parallel corpus of Manifesto de Prago with UD relations, to contribute to the development of research on Esperanto as a natural language. The distribution of dependency distances of Esperanto is not significantly different from the majority of the 21 languages in the corpus, thus partially supporting the claim that Esperanto is no less natural than other natural languages.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24K04089.

8. Bibliographical References

Yude Bi and Hua Tan. 2024. Language transfer in L2 academic writing: a dependency grammar approach. *Frontiers in Psychology*, 15:1–14.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Mark Fettes. 1996. *Manifesto de Prago*. Universala Esperanta Asocio.

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.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In W. O'Neil A.P. Marantz, A.P. Miyashita, editor, *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126. MIT Press.

Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Chris Gledhill. 2000. *The Grammar of Esperanto: A Corpus-based description*. Lincom.

Grant Goodall. 2023. Esperanto kaj lingvistiko: cent jaroj da (mal)amikeco. *Esperantologio / Esperanto Studies*, 4(12).

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial natural of linguistics input for sentential complexity. *Cognitive Science*, 29(2):261–290.

Richard Hudson. 1995. Measuring syntactic difficulty.

Jingyan 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.

Jingyang Jiang and Jinghui Ouyang. 2018. Minimization and probability distribution of dependency distance in the process of second language acquisition. In Jingyang Jiang and Haitao Liu, editors, *Quantitative Analysis of Dependency Structures*, pages 167–190. De Gruyter Mouton, Berlin, Boston.

Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2026. Changes in syntactic complexity indices with the language development of japanese as a second language: A longitudinal japanese learner corpus study. *Journal of Quantitative Linguistics*, 33(1):64–79.

Ilona Koutny. 2015. A typological description of esperanto as a natural language. *Język. Komunikacja.Informacja*, 10:43–62.

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.

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.

Haitao Liu. 2009. Probability distribution of dependencies based on chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3):256–273.

Haitao Liu. 2011. Quantitative analysis of Zamenhof's *Esenco kaj estonteco. Language Problems & Language Planning*, 35(1):57–81.

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.

Bill Manaris, Luca Pellicoro, George Pothering, and Harland Hodges. 2006. Investigating Esperanto's statistical proportions relative to other languages using neural networks and zipf's law. *Proceedings of the 24th IASTED International Conference on Artificial Intelligence and Applications*, pages 102–108.

Ruochen Niu, Yaquin Wang, and Haitao Liu. 2023. The cross-linguistic variations in dependency distance minimization and its potential explanations.

In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France.

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.

Jinghui Ouyang, Jingyang Jiang, and Haitao Liu. 2022. Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing*, 51:100603.

Masanori Oya. 2024. Cross-linguistic variances of dependency distances in multi-lingual parallel corpus. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 1166–1171.

Masanori Oya. 2025. UD treebanks for esperanto as a natural language. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 22–29.

Mikael Parkvall. 2000. How European is Esperanto?: A typological study. *Language Problems and Language Planning*, 34(1):63–79.

David Temberley. 2007. Minimization of dependency length in written english. *Cognition*, 105(2):300–333.

David Temperley. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Science*, 59:135–147.

Hengbin Yan and Yanghui Li. 2019. Investigating dependency distance across l2 modalities and proficiency levels. *Open Linguistics*, 5(1):601–614.

Lingyu Yi and Zhongqing He. 2026. Syntactic complexity in knowledge construction across disciplines: Evidence from dependency distance. *Journal of English for Academic Purposes*, 79:1475–1585.

Ludoviko Lazaro Zamenhof. 1887. *Dr. Esperanto's International Tongue*. Chaim Kelter.

Ludoviko Lazaro Zamenhof. 1907. *Esenco kaj Estonteco de la Ideo de Lingvo Internacia*. Hachette.

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.