Exploring Sampling Strategies for Linguistic Diversity: A Comparative Analysis of UD Treebanks

Althea Löfgren, Santiago Herrera, Sylvain Kahane, Bruno Guillaume, Natalia Levshina

Exploring Sampling Strategies for Linguistic Diversity:

A Comparative Analysis of UD Treebanks

Althea Löfgren[1], Santiago Herrera[2], Sylvain Kahane[3], Bruno Guillaume[4], Natalia Levshina[5]

Quantitative typology has experienced great advances in recent years, showing new ways of studying cross-linguistic variation and typological universals through the use of continuous data and corpus-based studies (*e.g.* Levshina (2019); Gerdes *et al.* (2021)). Universal Dependencies (UD) (de Marneffe *et al.* 2021), a syntax dependency framework and collection of treebanks, has proven to be an important resource for working with syntax in the field. However, the use of UD treebanks, like other corpora, must be considered with caution when typological goals are involved. While some works have pointed out intra-language and cross-linguistic variation due to annotation inconsistencies (Sinnemäki & Haakana 2020; Choi *et al.* 2021), sampling has received little attention.

Some studies acknowledge their results as preliminary, warning against drawing definitive conclusions (Guzmán Naranjo & Becker 2018; Gerdes *et al.* 2021). Some other works omit discussion of sampling altogether (Gerdes *et al.* 2019; Levshina 2019; Kahane *et al.* 2023). Given that cross-linguistic research is a key objective of UD, a sample with diverse genealogical coverage is essential for extrapolating cross-linguistic patterns and linguistic universals. While it is desirable to utilize the UD collection for research despite its limitations, moving forward prioritizing sampling is imperative for progress.

Of the 161 languages in the UD collection (v2.14), 75 belong to the Indo-European family. In comparison, there is a scarcity of languages from Africa (10), Australia (1), Papuanesia (4), and North America (3) . In the extensive literature on sampling, a common approach is to include at least one language from each genus or top-level family (Dryer 1989; Rijkhoff & Bakker 1998; Bickel 2008; Miestamo *et al.* 2016). Regardless of the classification scheme employed, UD falls short in this regard, encompassing languages from only 31 genera out of 246 (or 430 counting isolates; Hammarström *et al.* 2020) , including three isolates and one sign language.

From a quantitative typology point of view, high variability in corpus size makes good sampling difficult. Many languages have less than 1k tokens, while a decent description of the language could be achieved starting from 20k tokens. The size of Indo-European languages, as expected, is off-dimension, with German as the extreme example having more that 3810k tokens. In total, 11 out of the 31 genera represented in UD have less than 10k tokens.

The study aims to evaluate UD treebanks for typological analysis: assessing corpus size, representativeness, and linguistic biases. A case study on word order phenomena will be conducted (e.g. VO-order vs. noun-adjective order), employing various sampling techniques and phylogenetic regression to find word order correlations (Bakker 1998; Dryer 1989). The results are compared to the word order distributions in WALS to further evaluate the usefulness of UD in typological research. Since the sample has an Indo-European bias, we will also test on a sample of only Indo-European languages and compare with the global sample. The comparison of results from different sampling methods will hopefully offer some insights into the state of UD treebanks and how to best utilize them for typological research.

References

Bakker, Dik. 1998. *Flexibility and consistency in word order patterns in the languages of Europe.* De Gruyter Mouton. Pages 383–420.

Bickel, Balthasar. 2008. A Refined Sampling Procedure for Genealogical Control. *Language Typology and Universals*, **61**(08), 221–233.

Choi, Hee-Soo, Guillaume, Bruno, Fort, Karën, & Perrier, Guy. 2021. Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. *Pages 281–290 of:* Mitkov, Ruslan, & Angelova, Galia (eds), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021).* Held Online: INCOMA Ltd.

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

Dryer, Matthew. 1989. Large linguistic areas and language sampling. *Studies in Language*, **13**, 257–292.

Gerdes, Kim, Kahane, Sylvain, & Chen, Xinying. 2019. Rediscovering Greenberg's Word Order Universals in UD. *Pages 124–131 of:* Rademaker, Alexandre, & Tyers, Francis (eds), *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019).* Paris, France: Association for Computational Linguistics.

Gerdes, Kim, Kahane, Sylvain, & Chen, Xinying. 2021. Typometrics From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, **6**(02).

Guzmán Naranjo, Matías, & Becker, Laura. 2018 (12). Quantitative word order typology with UD. *In: Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018).*

Hammarström, Harald, Forkel, Robert, Haspelmath, Martin, & Bank, Sebastian. 2020. *Glottolog 4.2.1.* Max Planck Institute for the Science of Human History.

Kahane, Sylvain, Peng, Ziqian, & Gerdes, Kim. 2023. Word order flexibility: a typometric study. *Pages 68–80 of:* Rambow, Owen, & Lareau, François (eds), *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023).* Washington, D.C.: Association for Computational Linguistics.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, **23**(3), 533–572.

Miestamo, Matti, Bakker, Dik, & Arppe, Antti. 2016. Sampling for variety. *Linguistic Typology*, **20**(2), 233–296.

Rijkhoff, Jan, & Bakker, Dik. 1998. Language sampling. *Linguistic Typology*, **2**(01), 263–314.

Sinnemäki, Kaius, & Haakana, Viljami. 2020. Variation in Universal Dependencies annotation: A token-based typological case study on adpossessive constructions. *Pages 158–167 of:* de Marneffe, Marie-Catherine, de Lhoneux, Miryam, Nivre, Joakim, & Schuster, Sebastian (eds), *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020).* Barcelona, Spain (Online): Association for Computational Linguistics.

Notes

[1]Modyco, Université Paris Nanterre
[2]Modyco, Université Paris Nanterre
[3]Modyco, Université Paris Nanterre, CNRS, IUF
[4]Loria, Université de Lorraine
[5]Department of Language and Communication, Radboud University