Translationese in the Parallel Bible Corpus: Evaluating Extracted Word Order Features from Translated Texts

Typological databases such as WALS (Dryer and Haspelmath, 2013) and GramBank (Skirgård et al., 2023) typically present a categorical, often binary view of linguistic variation. Dividing languages into discrete typological categories invariably involves some degree of data reduction, which is problematic for features which exhibit non-bimodal distributions (Wälchli, 2009). Recent work (such as Levshina et al., 2023) has argued that a shift to gradient representation of such features would provide a more informative and accurate picture of actual cross- and intra-linguistic variation. Continuous representations are also preferrable when using typological data to inform multilingual language models and for other typologically informed NLP applications (Ponti et al., 2019).

Both Ponti et al. (2019) and Levshina et al. (2023) suggest extracting these gradient representations from parallel texts. The largest presently available parallel texts, such as the parallel Bible corpus (Mayer and Cysouw, 2014), contain text in a greater number of languages than are represented for many features in typological databases. Sentence- and word-level alignments also allow for counts of specific constructions to be computed automatically. Östling and Kurfalı (2023) apply this approach to data from the parallel Bible corpus in 1295 languages, and calculate token-level statistics for a number of syntactic features. The resulting gradient representations display a high degree of agreement with WALS data (when binarized), and capture a greater degree of intra-linguistic variation than the corresponding binary WALS features.

An important caveat of working with parallel texts like the Bible corpus is their translational nature, and in turn the potential effects of translational artefacts or "translationese" (Gellerstam, 1986). In addition to some general lexical and syntactic properties particular to translated texts, source language interference can be strong enough that source language phylogeny may be reconstructed just from a monolingual corpus of translated texts with different source languages (Rabinovich et al., 2017). Although Levshina et al. (2023) do not find any prominent impact of translationese when comparing gradient word order features extracted from translated texts in a parallel corpus to those extracted from original texts, they highlight the need for further investigation specifically for translations into low-resource languages.

We therefore aim to conduct an analysis of translationese effects on automatically extracted gradient word order features in as many languages as possible, exploiting the uniquely broad typological coverage of the parallel Bible corpus.

In our first approach, we apply Levshina et al.'s (2023) comparative method to a number of word order features, for all languages with sufficient data available in both the Bible corpus and *Universal Dependencies* (Nivre et al., 2020) treebanks. Preliminary findings show that for most analyzed dependencies, the counts extracted from Bible texts through annotation projection generally align well with those extracted from original UD texts.

As this method relies on dependency annotated original texts (which only exist for relatively few languages), a second approach is employed to investigate source language interference in a broader language sample, making use of source-and-translation text pairs in the Bible corpus. For a number of word order features, automatic document- and verse-level comparisons are made between each analyzed Bible translation and its respective source text; unexpectedly high levels of agreement in extracted word order preference between a given translation and its source text in a typologically distant language could indicate the presence of source language artefacts in the translated text.

References

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, pages 88–95. CWK Gleerup, Lund.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank Reveals Global Patterns in the Structural Diversity of the World's Languages. *Science Advances*, 9(16).

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, 13(1).

Robert Östling and Murathan Kurfalı. 2023. Language Embeddings Sometimes Contain Typological Generalizations. *Computational Linguistics*, pages 1–49.