Massively Multilingual Token-Based Typology Using the Parallel Bible Corpus

Amanda Kann
Stockholm University, Sweden
amanda.kann@su.se

Abstract

The parallel Bible corpus is a uniquely broad multilingual resource, covering over 1400 languages. While this data is potentially highly useful for extending language coverage in both token-based typology research and various low-resource NLP applications, the restricted register and translational nature of the Bible texts has raised concerns as to whether they are sufficiently representative of language use outside of their specific context. In this paper, we analyze the reliability and generalisability of word order statistics extracted from the Bible corpus from two angles: stability across different translations in the same language, and comparability with Universal Dependencies corpora and typological database classifications from URIEL and Grambank. We find that variation between same-language translations is generally low and that agreement with other data sources and previous work is generally high, suggesting that the impact of issues specific to massively parallel texts is smaller than previously posited.

Keywords: massively parallel texts, token-based typology, word order

1. Introduction

Typological databases are a valuable source of structured multilingual data. They are a vital resource for quantitative research in linguistic typology, but have also been used to inform cross-lingual transfer in multilingual language models (Philippy et al., 2023) and to evaluate to what extent these models extrapolate typological relationships directly from training data (Bjerva and Augenstein, 2021).

In describing cross-linguistic variation, typological databases like WALS (Dryer and Haspelmath, 2013) generally rely on sorting languages into discrete, often binary categories. This can result in the often misleading presentation of linguistic variation as bimodal (Wälchli, 2009), and the data reduction necessary to fit gradient variation into binary categories makes the resulting data less conceptually suitable for machine learning applications (Ponti et al., 2019).

An alternative approach is to infer typological properties directly from text, producing a more granular representation which better captures intra-linguistic variation. Levshina (2019) terms this approach *token-based typology*, and uses Universal Dependencies corpora to investigate word order variability in 60 languages.

A disadvantage of corpus-based approaches to typology is that they generally rely on annotated data. This requirement is at odds with the desire for high language coverage, as the availability of annotated corpora and automatic parsing tools is limited and skewed towards already well-studied languages – an imbalance which remains highly present in all areas of NLP research (Joshi et al., 2020).

The parallel nature of massively parallel texts (Cysouw and Wälchli, 2007) can be leveraged to ameliorate this problem: through cross-lingual word alignment, annotations can be projected from a small subset of parsed corpora to the entire parallel corpus (Östling, 2015). Östling and Kurfalı (2023) use word alignment and projection of dependency relations in a massively parallel corpus of Bible translations to produce token-level word order statistics in 1295 languages, displaying high levels of agreement with typological databases for most evaluated word order features.

This Bible corpus (Mayer and Cysouw, 2014) is the "most parallel" text currently available by a considerable margin, containing 1846 translations in 1401 languages (in the version used by Östling and Kurfalı 2023). Beyond its applications for linguistic typology, it has been used as training data for low-resource PoS tagging (Imani et al., 2022) and to improve the zero-shot cross-lingual transfer performance of pretrained multilingual language models (Ebrahimi and Kann, 2021).

The use of massively parallel texts such as the Bible corpus for token-based typology is not in itself unproblematic, however. These texts generally only represent a single domain-specific and, in the case of the Bible corpus, typically archaic doculect rather than a balanced sample of registers and language varieties (Levshina, 2022). Parallel texts are also by necessity translational, which is another potential cause of dissimilarity to original texts. The impact of these issues on the generalisability of conclusions drawn from domain-specific parallel data seems to depend on the type of linguistic property in focus – frequencies of specific lexical items, for instance, are likely to vary

LREC-COLING 2024, pages 11070–11079

Word order feature	Dependency definition	URIEL feature	Grambank feature
Adjective/noun order	ADJ <-amod- NOUN	S_ADJECTIVE_AFTER_NOUN	GB193
Adposition/noun order	ADP <-case- NOUN	S_ADPOSITION_AFTER_NOUN	GB074 + GB075
Numeral/noun order	NUM <-nummod- NOUN	S_NUMERAL_AFTER_NOUN	GB024
Object/verb order	NOUN/PROPN <-obj- VERB	S_OBJECT_AFTER_VERB	–
Oblique/verb order	NOUN/PROPN <-obl- VERB	S_OBLIQUE_AFTER_VERB	–
Relative/noun order	VERB <-acl- NOUN	S_RELATIVE_AFTER_NOUN	GB327 + GB328
Subject/verb order	NOUN/PROPN <-nsubj- VERB	S_SUBJECT_AFTER_VERB	–

Table 1: Labels and definitions of examined word order features, taken from Östling and Kurfalı 2023 and appended with binarized Grambank feature combinations (where corresponding features exist in Grambank).

greatly between text types, while word order statistics seem to be relatively stable across corpora of different domains and sizes (Levshina, 2019; Choi et al., 2021). However, it has not yet been examined whether this property holds for the specific domain of the Bible corpus, nor if translational artefacts have a significant impact on the reliability of word order statistics inferred from parallel texts.

This paper therefore aims to examine to what degree the Bible corpus (and, by extension, similar massively parallel texts) can be useful for token-based word order typology through two approaches:

(i) analyzing the stability of extracted word order statistics across different Bible translations in the same language;

(ii) a three-way cross-linguistic comparison between word order statistics from Bible data, comparable statistics from reference corpora in other domains and classifications from typological databases.

2. Data

2.1. Bible corpus data

Word order feature statistics from the Bible corpus were obtained from the *Parallel text typology* dataset (Östling and Kurfalı, 2023). The dataset contains statistics for 7 word order features computed from projected PoS and dependency annotations across 1664 doculects[1] in 1295 languages, although not all features are available for each doculect. A list of these features is provided in Table 1. The word order features are binarized – for a pair of constituents (such as adjective and noun) with a given dependency relation (amod), the dependent constituent can either precede (ADJ-N) or follow (N-ADJ) its head. For each doculect and dependency relation (for which there existed sufficient data to compute this statistic), a value between 0 and 1 represents the proportion of head-dependent order among occurrences of the constituent pair. For example, an adjective/noun order value of 0.8 corresponds to the constituent order being N-ADJ in 80% of amod dependency relation occurrences (and ADJ-N in the remaining 20%). For a thorough overview of how these statistics were computed, including an evaluation of the dataset's alignment with other typological data sources, please see Östling and Kurfalı 2023.

[1]The term *doculect* is henceforth used to refer to the language contained in a single Bible translation, in alignment with Östling and Kurfalı 2023. The parallel Bible corpus may contain multiple translations of the Bible (and hence multiple doculects) in a given single language.

2.2. Classifications from typological databases

We used data from two manually assembled typological databases: (i) URIEL (Littell et al., 2017), which aggregates features from WALS (Dryer and Haspelmath, 2013) and Ethnologue (Eberhard et al., 2022), and (ii) Grambank (Skirgård et al., 2023). These databases contain binary (or, in the case of some Grambank features, trinary) classifications for various typological features in a large number of languages. Although there is no direct 1-to-1 correspondence between features in URIEL and Grambank, comparable combinations of features were available for 4 of the 7 examined word order features (see mapping between data sources in Table 1). In some cases, such as for numeral/noun order in Galo (adl, Sino-Tibetan), there was disagreement between classifications in URIEL (N-NUM) and Grambank (NUM-N). In this specific case, the projected Bible word order agreed with URIEL – examination of the two primary data sources used for this feature-language pair in each database revealed that N-NUM indeed seems to be the preferred order, and that the Grambank classification is likely erroneous. This suggests in itself that extracted token-level statistics can be useful for informing judgments about conflicts between typological databases, in line with previous work (Choi et al., 2021; Östling and Kurfalı, 2023).

2.3. UD reference data

Reference corpora were obtained from version 2.12 of Universal Dependencies (Zeman and Nivre, 2023), consisting of 245 treebanks in 141 languages, of which 79 were also present in the Bible dataset.

Following Choi et al. 2021, we used the *GREW* tool (Guillaume, 2021) to extract occurrences of the PoS tags and dependency relations corresponding to each word order feature listed in Table 1, and computed per-language word order statistics in the manner described in section 2.1.[2]

3. Variation across same-language doculects

For 159 of the 1295 languages included in the *Parallel text typology* dataset, word order statistics were available for more than one doculect. Just like different-language translations, doculects in the same language may differ across a number of properties – source text, translator, purpose or *skopos* of the translation (De Vries, 2007), age, set of translated verses[3], etc. – which may all be sources of variation when computing token-based statistics from these texts.

To estimate the degree of inter-doculect variation caused by these properties for a given word order feature f, we first established a subset N_f from the 159 multiple-doculect language set N, excluding languages in which the word order statistic corresponding to that feature was not available across all doculects. For each language L in N_f, we computed the standard deviation σ_L of the relevant word order statistic across all doculects. We then calculated the mean $\mu(\{\sigma_L\})$ and dispersion $\sigma(\{\sigma_L\})$ across all languages in N_f. Since the bounds of the word order statistic are 0 and 1, a mean $\mu(\{\sigma_L\})$ of 0 would correspond to complete inter-doculect stability (that is, no variation between same-language doculects in the proportions of observed word order counts) for a given feature. Based on previous studies of intra-language (but cross-domain) word order variation (Levshina, 2019; Choi et al., 2021), we expect an overall high degree of stability between same-language doculects.

The results of this analysis for each word order feature in the Bible dataset are provided in Table 2. As expected, inter-doculect variation was consistently low across all features; the difference between mean standard deviation for the most (Num/N order) and least (Adj/N order) stable features was 0.042. This suggests that the high degree of consistency observed by Levshina (2019) and Choi et al. (2021) also holds within the Bible text domain for a variety of word order features.

Word order feature	$\lvert N_f \rvert$	$\mu(\{\sigma_L\})$	$\sigma(\{\sigma_L\})$
Adjective/noun order	157	0.065	0.083
Adposition/noun order	101	0.046	0.072
Numeral/noun order	157	0.023	0.029
Object/verb order	152	0.032	0.047
Oblique/verb order	159	0.039	0.036
Relative/noun order	134	0.044	0.045
Subject/verb order	158	0.042	0.044

Table 2: Mean and standard deviation of inter-doculect variation (standard deviation of the word order statistics of all doculects in a given language) across all languages, for each word order feature.

Some variation in stability across languages was observed within each feature, however, as indicated by the relatively high dispersion values in Table 2. Although inter-doculect variation for a given feature was low (between 0.0 and 0.05) for a majority of languages, a small number of languages had large differences in word order proportions between doculects. Grouping the Adj/N analysis by the classification of each language in typological databases reveals that variation is considerably higher among languages categorised as "both orders occur" than among languages categorised as preferring one order over the other, as shown in Figure 1. This analysis also reveals a number of outlier languages among Adj-N and N-Adj languages. The farthest outlier in the Adj-N category, Sango (sag, Atlantic-Congo), has two

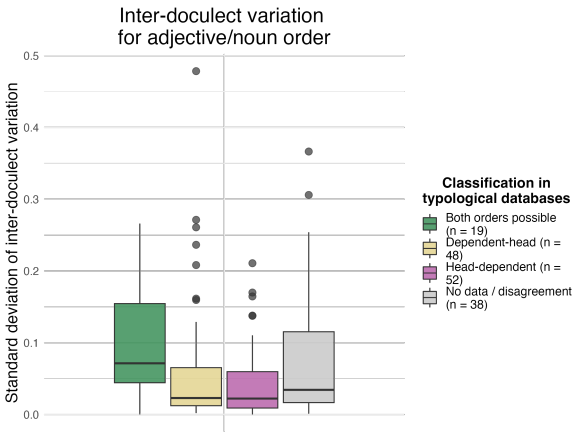

Figure 1: Inter-doculect variation per language for adjective/noun order, grouped by classification according to typological databases.

[2]No statistic was computed for dependency relations which occurred fewer than 20 times in a given language, following Levshina 2019.

[3]Partial translations (where fewer than 80% of New Testament verses are translated) were excluded in the creation of the dataset (Östling and Kurfalı, 2023).

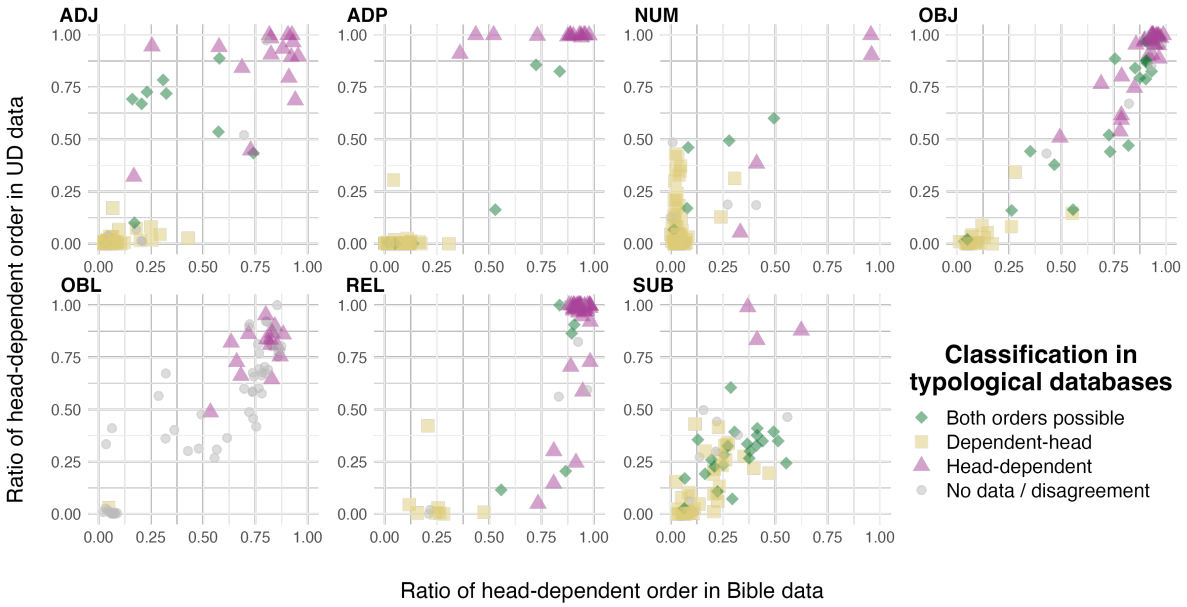

Figure 2: Three-way comparison between projected Bible data, manually annotated UD corpora and typological database classifications. The individual graph labels correspond to the first three letters of the corresponding word order feature's URIEL feature label (see Table 1).

doculects; one agrees with the ADJ-N classification, while the other has predominantly N-ADJ order. Sango is primarily an oral language, and is spoken alongside French (a predominantly N-ADJ language) in the Central African Republic (Thornell, 1995) – influence from French may explain the high degree of variation between the doculects, although deeper analysis is needed. As evidenced by this example, investigating inter-doculect variation for a given feature seems to be a useful method for identifying outlier cases in token-based typology studies and multilingual NLP applications which rely on the parallel Bible corpus.

4. Three-way comparison

To investigate to what degree the representativity issues raised by Levshina (2022) and others may affect conclusions about word order typology drawn from the Bible data, we compare the word order feature statistics described in section 2.1 to similarly extracted statistics from UD data of a variety of domains and text types, along with typological classifications from URIEL and Grambank.

The results of this comparison are visualized in Figure 2. Across all features, the three datasets displayed high levels of agreement – in most cases, feature-language pairs classified as either head-dependent or dependent-head had correspondingly high or low proportions of head-dependent order occurrences in both the Bible data and the UD reference data. The object/verb order results provide a particularly clear example of Bible-UD agreement – even for most languages where both VO and OV orders are possible, the proportion of occurrences of the two types was highly similar in Bible and UD data. This result mirrors Levshina 2019's findings from token-based analysis of UD corpora, where object/verb order was found to have high cross-linguistic variability but low intra-linguistic variability, and suggests that such conclusions could also be reliably drawn from domain-specific data such as the Bible corpus.

The results for subject/verb order are also in line with Levshina 2019's findings of low variability both across and within languages: in both Bible and UD data, languages in which both SV and VS order can occur almost always had a higher proportion of SV occurrences. The emergence of this pattern is an interesting result of token-based methods, which would have been considerably more difficult to find with a less granular metric of word order preference. For numeral/noun order, the two N-NUM classed languages which have a high proportion of NUM-N occurrences in both Bible and UD data – Korean (`kor`, Koreanic) and Thai (`tha`, Tai-Kadai) – also seem to be cases of genuine variation which is not captured by the binary classifications of typological databases.

Additional points of slight disagreement for the numeral/noun feature are found among many languages classified as NUM-N, for which more N-NUM occurrences were found in UD data than in Bible data. This discrepancy (and similar patterns

for other word order features) has a number of potential explanations – varying distributions of lexical items in the texts' different domains, or translator preference for a particular order resulting in translational bias in the Bible data – which warrant further investigation. Another potential cause is serial word order variation: a considerable number of languages have a different NUM/N order for the cardinal numeral 1, which is often used as an indefinite article or for another non-enumerative function as well, than for all other cardinal numerals (Kann, 2019). This is taken into account by Östling and Kurfalı (2023) who limit their analysis to the numerals 2–9, while the partial absence of cross-lingually consistent lexical annotation in the UD corpora did not allow for this type of narrow-scope analysis in the present study. Following Östling and Kurfalı 2023, restricting comparisons to specific semantic concepts or syntactic constructions for all relevant word order features would increase comparability further and likely yield interesting results; this is a promising direction for future work.

Finally, as Levshina (2019) notes, numeral/noun order is an areally divergent feature, meaning that an unstratified language sample can skew results for this feature. Although the restricted overlap between languages in UD and Bible data limits the coverage of this three-way analysis, it would be interesting to compare data only between the Bible corpus and typological databases with a broader and more balanced language sample for this feature in particular.

5. Conclusions

By investigating the inter-doculect stability and cross-resource comparability of word order statistics extracted from annotation projections in the parallel Bible corpus, we have addressed issues of reliability and generalisability concerning the use of Bible texts and similar massively parallel texts for token-based typology.

For both analyses, our results were generally in line with prior token-based work using multi-domain UD corpora, indicating that word order statistics extracted from massively parallel texts (even with a restricted domain and a lack of manual annotation) can often be sufficiently representative of broader language use. We hope that this finding, along with its caveats discussed in the paper, will be of use for extending language diversity and coverage both in token-based typology and in NLP applications using parallel texts.

Finally, we have outlined some directions for future work surrounding the parallel Bible corpus, such as investigating lexically dependent word order variation and potential translational artefacts with a larger, stratified language sample – there are still many facets of this uniquely broad language resource which have not yet been explored.

6. Limitations and ethical considerations

The chosen methods' reliance on either multiple doculects or supplementary data for each analyzed language meant that only a relatively small subset of languages in the Bible dataset could be analyzed, leading to low language coverage compared to using the entire Bible dataset. This also naturally skewed the language sample towards higher-resource languages, introducing further bias. The areal and phylogenetic distribution of the language sample is described in Table 3 and Table 4 in the appendix, and full lists of languages included in the study are provided as supplementary material to the paper.

Only a limited set of word order features were analyzed in this study. It is not certain that the same methods would yield similar results for other syntactic properties, or features on another linguistic level altogether, and the results should not be extrapolated across linguistic domains without further feature-specific investigation.

Finally, it is important to emphasize that the results of this paper do not suggest that it is advisable to rely exclusively on Bible texts for either quantitative typology or low-resource NLP applications. Both the culturally specific nature and history of the text itself and the scarcity of metadata concerning the translation process (including the involvement of native speakers) for most doculects in the Bible corpus should be taken into account in extraction of anything other than general structural properties of the text.

7. Acknowledgments

This work was made possible by individual PhD student funding from the Department of Linguistics at Stockholm University. We are especially grateful to Robert Östling, Bernhard Wälchli and Viggo Kann for providing invaluable input on this paper as it developed.

8. Data/code availability statement

All data related to the parallel Bible corpus used in this paper can be found in v1.0.0 of the *Parallel text typology* dataset (Östling and Kurfalı, 2023), released under the CC BY-SA 4.0 license.

The versions of all reference datasets and databases used in this paper are freely available online under open licences – see the Language Resource References section for persistent identifiers.

All code used to produce the results detailed in this paper is available at https://doi.org/10.5281/zenodo.10858784, under the GPL-3.0 license.

9. Bibliographical References

Johannes Bjerva and Isabelle Augenstein. 2021. Does Typological Blinding Impede Cross-Lingual Sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Hee-Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating dominant word order on Universal Dependencies with graph rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 281–290, Held Online. INCOMA Ltd.

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*, 60(2):95–99.

Lourens De Vries. 2007. Some remarks on the use of Bible translations as parallel texts in linguistic research. *Language Typology and Universals*, 60(2):148–157.

Abteen Ebrahimi and Katharina Kann. 2021. How to Adapt Your Pretrained Multilingual Model to 1600 Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, and Sebastian Bank. 2024. Glottolog 5.0.

Ayyoob Imani, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-Based Multilingual Label Propagation for Low-Resource Part-of-Speech Tagging. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1577–1589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amanda Kann. 2019. *Ordföljdsvariation inom kardinaltalssystem: Extraktion av ordföljdstypologi ur parallella texter*. Bachelor's thesis, Stockholm University, Stockholm.

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

Natalia Levshina. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

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 (LREC'20)*, pages 4034–4043, Marseille, France. European Language Resources Association (ELRA).

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

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.

Christina Thornell. 1995. Established French loanwords in Sango - A pilot study. Technical Report 44, Department of Linguistics, Lund University.

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, 13(1):77–94.

Robert Östling. 2015. Word Order Typology through Multilingual Word Alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Robert Östling and Murathan Kurfalı. 2023. Language Embeddings Sometimes Contain Typological Generalizations. *Computational Linguistics*, 49(4):1003–1051.

10. Language Resource References

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2022. *Ethnologue: Languages of the World*, 25th edition. SIL International, Dallas, Texas.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Lennart Chevallier, Swintha Danielsen, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Giada Falcone, Anna-Maria Fehn, 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 H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Alexandra Marley, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya, Michael Müller, Saliha Muradoglu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Jesse Peacock, India O. C. Pearey, Naomi Peck, Jana Peter, 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, Dineke Schokkin, Jeff Siegel, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, 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 v1.0 (v1.0.3).

Daniel Zeman and Joakim Nivre. 2023. Universal Dependencies 2.12.

Robert Östling and Murathan Kurfalı. 2023. Parallel text typology dataset (1.0.0).

Appendices

A – Distribution of languages in inter-doculect variation study

Table 3 presents an overview of the number of languages included in the inter-doculect variation study for each word order feature, grouped by macroarea and language family classification according to Glottolog (Hammarström et al., 2024).

B – Distribution of languages in three-way comparison study

Table 4 provides a similar languages-per-feature overview for the three-way comparison study, grouped by macroarea and language family classification according to Glottolog (Hammarström et al., 2024). The feature classification of these languages according to data from URIEL and Grambank is presented in Table 5.

Glottolog classification	Number of studied languages per feature						
Macroarea	**ADJ**	**ADP**	**NUM**	**OBJ**	**OBL**	**REL**	**SUB**
Eurasia	65	57	64	65	65	64	65
Africa	31	19	31	30	31	26	31
Papunesia	29	14	29	25	30	21	30
North America	18	4	18	18	18	13	18
South America	14	7	15	14	15	10	14
Family							
Indo-European	43	39	43	43	43	43	43
Atlantic-Congo	23	14	23	23	23	21	23
Austronesian	19	14	19	19	19	18	19
Mayan	13	3	13	13	13	8	13
Turkic	10	9	10	10	10	10	10
Afro-Asiatic	6	5	6	6	6	5	6
Nuclear Trans New Guinea	4	0	4	2	5	1	5
Uto-Aztecan	3	0	3	3	3	3	3
Uralic	3	3	3	3	3	3	3
Tupian	3	3	3	3	3	3	3
Quechuan	3	2	3	3	3	3	3
Sino-Tibetan	2	0	1	2	2	2	2
Isolate	2	1	2	2	2	1	2
Hmong-Mien	2	0	2	2	2	2	2
Dravidian	2	2	2	2	2	2	2
Sepik	1	0	1	0	1	0	1
Uru-Chipaya	1	0	1	1	1	1	1
Arawakan	1	0	1	1	1	0	1
Austroasiatic	1	1	1	1	1	1	1
Aymaran	1	0	1	1	1	1	1
Tucanoan	1	1	2	2	2	0	1
Tai-Kadai	1	1	1	1	1	1	1
Ta-Ne-Omotic	1	1	1	1	1	0	1
Border	1	0	1	1	1	1	1
Chicham	1	0	1	1	1	0	1
Koreanic	1	1	1	1	1	1	1
Pidgin	1	1	1	1	1	1	1
Otomanguean	1	0	1	1	1	1	1
Chiquitano	1	0	1	1	1	1	1
Nuclear Torricelli	1	0	1	1	1	0	1
Nilotic	1	0	1	1	1	0	1
Chocoan	1	0	1	0	1	0	1
Mande	1	0	1	0	1	0	1
Ndu	1	0	1	0	1	0	1

Table 3: Macroarea and family distribution of languages included in the inter-doculect variation study described in section 3.

Glottolog classification	Number of studied languages per feature						
Macroarea	**ADJ**	**ADP**	**NUM**	**OBJ**	**OBL**	**REL**	**SUB**
Eurasia	58	54	56	60	57	52	59
Africa	6	7	7	7	6	4	7
Papunesia	3	4	2	4	4	2	4
North America	2	1	2	2	2	2	2
South America	1	1	1	5	4	1	4
Australia	0	0	0	1	0	0	1
Family							
Indo-European	41	39	38	42	38	36	41
Turkic	6	4	6	6	6	4	6
Afro-Asiatic	4	5	4	4	4	3	4
Uralic	3	3	4	4	4	4	4
Austronesian	3	4	2	4	4	2	4
Dravidian	3	3	3	3	3	2	3
Mayan	1	1	1	1	1	1	1
Tupian	1	1	1	3	3	1	2
Tai-Kadai	1	1	1	1	1	1	1
Mongolic-Khitan	1	0	1	1	1	1	1
Mande	1	1	1	1	1	0	1
Koreanic	1	1	1	1	1	1	1
Isolate	1	1	1	1	1	1	1
Austroasiatic	1	1	1	1	1	1	1
Atlantic-Congo	1	2	2	2	2	2	2
Uto-Aztecan	1	0	1	1	1	1	1
Arawakan	0	0	0	1	0	0	1
Nuclear-Macro-Je	0	0	0	1	1	0	1
Pama-Nyungan	0	0	0	1	0	0	1

Table 4: Macroarea and family distribution of languages included in the three-way comparison study described in section 4.

Word order feature	**Total**	Both occur	Dep-head	Head-dep	No data/disagreement
Adjective/noun order	**70**	10	36	16	8
Adposition/noun order	**67**	8	38	14	7
Numeral/noun order	**68**	5	53	4	6
Object/verb order	**79**	20	19	32	8
Oblique/verb order	**73**	0	1	15	57
Relative/noun order	**61**	6	8	38	9
Subject/verb order	**77**	23	43	3	8

Table 5: Number of languages included in the three-way comparison study described in section 4 for each word order feature, grouped by classification in typological databases.