Grambank's typological advances support computational research on diverse languages

Hannah J. Haynie
University of Colorado Boulder
Boulder, Colorado, USA
hannah.haynie@colorado.edu

Damián Blasi
Harvard University
Cambridge, Massachusetts, USA
damianblasi@gmail.com

Hedvig Skirgård
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
hedvig_skirgard@eva.mpg.de

Simon J. Greenhill
University of Auckland
Auckland, New Zealand
simon.greenhill@auckland.ac.nz

Quentin D. Atkinson
University of Auckland
Auckland, New Zealand
q.atkinson@auckland.ac.nz

Russell D. Gray
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
russell_gray@eva.mpg.de

Abstract

In spite of increasing attention on less-resourced languages in Natural Language Processing (NLP), equitable access to language technologies and inclusion of diverse languages in the development of these technologies remains a problem (Joshi et al., 2020). This disparity in resources and research attention is pronounced – only a handful of the world's approximately 7,000 languages receive the majority of scholarly attention (Blasi et al., 2022). Extending the reach of language technologies to diverse, less-resourced languages is important for tackling the challenges of digital equity and inclusion, and incorporating typological information into language transfer and multilingual learning is an important strategy for doing this. Here we introduce the Grambank typological database as a resource to support efforts that leverage typological features to enhance multilingual NLP.

To date, the cross-linguistic information about morphology and syntax that has been recruited for NLP comes primarily from datasets designed for theoretical linguistics research, with very little consideration of how this data may be used in computational tasks (Dryer and Haspelmath, 2013; Michaelis et al., 2013; Bickel and Nichols, 2002). As a result these existing typological datasets suffer from several limitations, including small numbers of adequately annotated languages, excessive missing data per feature, and lack of transparency in the content and coding of features (O'Horan et al., 2016). Grambank is a resource designed and curated by linguistic typologists to serve both theoretical linguistic purposes and computational uses. Its 195 morphosyntactic features cover a similar range of grammatical phenomena as prior typological databases (e.g. word order, grammatical relation marking, constructions like interrogatives and negation), but Grambank differs in its design in ways that facilitate its use in computational research.

Each of Grambank's features encodes some characteristic of the morphology and/or syntax of languages. The content of the feature set balances the description of a wide range of structures that are known to vary across languages with the availability of information for a maximal set of languages. Feature names take the form of a question (e.g. 'Are there prenominal articles?'), and values for a majority of features are binary (0/'no', 1/'yes'). Six word order features have multi-state values (e.g. 'Order A', 'Order B', or 'Both Order A and Order B'), which can easily be binarised for analytical purposes. Binary feature values avoid the ambiguity of binned or inadequately described categories, and the representation of Grambank datapoints in terms of the presence or absence of linguistic traits allows the dataset to report all strategies identified in empirical sources for expressing a particular meaning or function. This contrasts with prior typological resources that encode a single 'dominant' category per meaning or function (Dryer and Haspelmath, 2013).

The typological content of Grambank is structured as a simple list of features, with no hi-

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 147–149

erarchical relationships between features (e.g. specific characteristics that are only coded if a certain value is registered for a more general feature). Care was also taken to avoid strict logical dependencies between features (i.e. situations where a certain value for Feature A entails a particular value for Feature B). Functional dependencies may still exist between features for a variety of reasons, such as communicative pressures or common processes of language change. However, the structure of the dataset eliminates a great deal of the redundancy in typological data that is problematic for tasks such as measuring language distances (Hammarström and O'Connor, 2013).

To promote transparency (Slingerland et al., 2020), the Grambank web interface includes extensive documentation for each feature, including step-by-step procedures that outline the analytical decisions made by annotators in determining feature values, illustrative examples from languages with different feature values, and references to relevant theoretical literature.

Grambank is annotated by linguists based on descriptions (e.g. published grammars) of languages. It currently includes data for 2,467 languages – around a third of the world's total linguistic diversity – from 316 different language families around the globe. This sample covers all continents (Antarctica excepted), and all 24 linguistically relevant geographic areas identified in prior research (Nichols et al., 2013). Whereas NLP research to date features languages of continental Eurasia almost exclusively, only about 20% of the Grambank sample is drawn from this region, with the remainder representing diverse languages from Africa, the Americas, Australia, Papua New Guinea, and Oceania. While Grambank is not intended to be a perfect stratified sample of language families or macroareas, it provides representation of areas and languages that are often under-sampled, including minority languages, endangered languages, languages from small language families, and isolates.

The size of the language sample in Grambank is similar to WALS (Dryer and Haspelmath, 2013), but Grambank represents a tremendous leap forward in terms of the overall number of datapoints available for characterizing individual languages, investigating language universals and tendencies, and examining the full range of grammatical diversity. Grambank advances the field by making complete or nearly complete sets of high quality, easily interpreted grammatical information available for a large and diverse set of languages. Missing data has been repeatedly presented as the most important limitation of typological data for use in multilingual NLP (O'Horan et al., 2016; Ponti et al., 2019; Bjerva et al., 2020), and this is where Grambank most clearly exceeds the prior benchmark. On average a WALS feature is coded for approximately 400 languages (Dryer and Haspelmath, 2013). In contrast, Grambank features are coded for approximately 1,500 languages on average. This means that a typical language in WALS is coded for only approximately 30 features, while it is likely to be coded in Grambank for approximately 145 features. In sum, approximately 17% of the potential datapoints in WALS have values (Dryer and Haspelmath, 2013; O'Horan et al., 2016), while Grambank pushes the total completion rate above 70%.

Typological data has been shown to be a useful tool for improving the performance of multilingual methods (Zhang et al., 2012; Ammar et al., 2016), transfer of technologies from high resource languages (Naseem et al., 2012), and a variety of other tasks that enable multilingual NLP and ultimately the development of inclusive language technologies (Rama and Kolachina, 2012; Östling, 2015; Takamura et al., 2016). Grambank represents a significant advance in the typological information that can be used to support these activities.

Limitations

The resource described herein includes information for only approximately one third of the languages of the world; its use for computational tasks involves some risk of bias related to sampling based on availability of grammatical descriptions and risk of excluding understudied languages.

The evaluation of the Grambank resource presented here relies on qualitative differences between this resource and the existing state of the art in cross-linguistic morphosyntactic data. Further analyses are warranted to examine the impacts of this resource on specific tasks.

Ethics Statement

This research complies with the principles of the ACL Ethics Policy. Cross-linguistic morphosyntactic resources have the potential to aid in the expansion of computational resources to less-resourced languages, but we note that the needs and interests of language communities vary and that digital equity and inclusivity require the involvement of those communities in research and development of technologies.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics*, pages 33–40.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

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

Harald Hammarström and Loretta O'Connor. 2013. Dependency-sensitive typological distance. In Anju Saxena and Lars Borin, editors, *Approaches to Measuring Linguistic Differences*, pages 329–352. De Gruyter, Berlin.

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.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Johanna Nichols, Alena Witzlack-Makarevich, and Balthasar Bickel, editors. 2013. *The AUTOTYP genealogy and geography database: 2013 release*. University of Zürich.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

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.

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of COLING 2012: Posters*, pages 975–984, Mumbai, India. The COLING 2012 Organizing Committee.

Edward Slingerland, Quentin D. Atkinson, Carol R. Ember, Oliver Sheehan, Michael Muthukrishna, Joseph Bulbulia, and Russell D. Gray. 2020. Coding culture: Challenges and recommendations for comparative cultural databases. *Evolutionary Human Sciences*, 2:E29.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 69–76, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378, Jeju Island, Korea. Association for Computational Linguistics.