Multilingual Gradient Word-Order Typology from Universal Dependencies

Emi Baylor*
McGill University
Mila Quebec AI Institute
emily.baylor@mail.mcgill.ca

Esther Ploeger*
Dept. of Computer Science
Aalborg University
espl@cs.aau.dk

Johannes Bjerva
Dept. of Computer Science
Aalborg University
jbjerva@cs.aau.dk

Abstract

While information from the field of linguistic typology has the potential to improve performance on NLP tasks, reliable typological data is a prerequisite. Existing typological databases, including WALS and Grambank, suffer from inconsistencies primarily caused by their categorical format. Furthermore, typological categorisations by definition differ significantly from the continuous nature of phenomena, as found in natural language corpora. In this paper, we introduce a new seed dataset made up of continuous-valued data, rather than categorical data, that can better reflect the variability of language. While this initial dataset focuses on word-order typology, we also present the methodology used to create the dataset, which can be easily adapted to generate data for a broader set of features and languages.

1 Introduction

Data from the field of linguistic typology has the potential to be useful in training NLP models (Bender, 2016; Ponti et al., 2019). However, the main existing typological databases, WALS (World Atlas of Language Structures) (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023), contain inconsistent and contradictory information (Baylor et al., 2023). These issues stem, in large part, from the categorical format of the data, which is over-simplistic and therefore cannot capture the nuance and variability that exist in natural language.

For example, one of the features describes the ordering of adjectives and the noun they modify. The categories in these datasets are Noun-Adjective, Adjective-Noun, or Variable. Limiting the options to these three categories removes any information differentiating a language that employs Noun-Adjective ordering 10% of the time from one that does so 90% of the time. In addition, the threshold between the Noun-Adjective and Adjective-Noun categories and the Variable category is often not clear, which can lead to inconsistencies in the data. As an example, the same 90% Noun-Adjective language might be classified as Variable in one database, but might be seen as consistently Noun-Adjective enough to be classified in the Noun-Adjective category in another database.

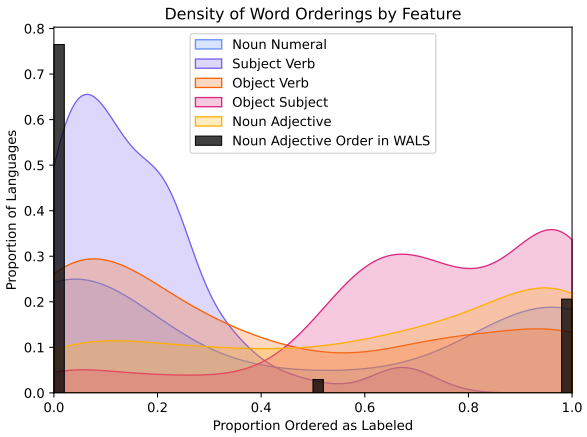

Figure 1: Proportion of languages with proportion of relevant words ordered as labeled, by feature. The black represents WALS Noun Adjective categories, with the far left being the Adjective Noun languages, the far right being the Noun Adjective languages, and the center being the variable languages. All other distributions come from our dataset.

In this paper, we apply recommendations presented in Levshina et al. (2023) and extend their analysis by introducing a new continuous-valued typological dataset that removes the need to oversimplify data into categories. In particular, we focus on word-level typology, and present a method for extracting gradient typology that utilizes the part of speech annotations available in the Universal Dependencies (UD) treebank corpus (Nivre et al., 2020). We then propose a novel regression-based typology task.

This new dataset and the methods used to create it are beneficial not only to NLP, but also potentially

* These authors contributed equally to this work.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 42–49

French phrase	Noun-Adjective Count	Adjective-Noun Count	English Translation
Mon *cher* ami	0	1	*"My dear friend"*
Mon appartement *ancien*	1	0	*"My ancient apartment"*

Table 1: An example of counting Noun-Adjective and Adjective-Noun instances in the dataset creation process, with English translations for ease. French nouns are underlined and in purple, and French adjectives are italicized and in *red*.

to the field of linguistic typology itself. Similar to previous works that include automatically recognizing or utilizing typological information (Asgari and Schütze, 2017; Saha Roy et al., 2014; Nikolaev and Pado, 2022), we introduce a new data collection technique that can methodically extract typological information from existing annotated text-based datasets.

2 Background

2.1 Linguistic Typology

Linguistic typology is the study of the world's languages through the comparison of specific features of language, across a variety of languages (Ponti et al., 2019; Comrie, 1988). These features can focus on any aspect of language, including phonology (Hyman, 2008; Lindblom and Maddieson, 1988), syntax (Greenberg, 1966; Comrie, 1989), morphology (Comrie, 1989), and phonetics (Lindblom and Maddieson, 1988).

For example, a typologist might look to contrast the number of distinct vowels that a diverse group of languages employ (Maddieson, 2013). Or they would compare how different languages tend to order verbs and their subjects: do verbs generally occur before or after their subjects in a sentence? (Dryer, 2013). Compared to other areas of linguistics, word order data can be relatively easy to collect, meaning that word order features tend to have data across a large number of the world's languages. Additionally, within NLP, word-order is the most commonly studied typological feature when it comes to cross-lingual transfer (Philippy et al., 2023). Typological diversity is furthermore used in NLP as an argument for language sampling, albeit without any consensus for the underlying meaning of the term (Ploeger et al., 2024).

2.2 Existing Typological Resources

The current two most popular typological databases, WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023), offer coverage of over 2,000 languages each. While the overall scope of the databases is great, their reliance on categorical representations of linguistic features means they frequently oversimplify data to the point of creating inconsistencies and errors, both within the databases, and with each other. Although this categorical distinction is a conscious design choice, we argue that a data driven and gradient solution can provide benefits both for typology and for NLP.

One solution to this problem of discrete categorical representations, proposed by Levshina et al. (2023), is to instead replace them with gradient representations. These continuous gradient representations are better able to capture nuanced linguistic information.

3 Continuous-Valued Seed Dataset

We introduce a seed dataset based on the idea of continuous representations of linguistic features (Levshina et al., 2023). This dataset is currently small, with coverage of fewer than 100 languages across a limited number of features. However, the process used to create it, described in section 3.1, can be easily adapted for broader feature coverage, as well as broader language coverage.

3.1 Dataset Creation

To best describe the creation of this dataset, we will walk through the data collection process for a single linguistic feature: the relative orderings of adjectives and the nouns they modify. In WALS (feature 87A) and Grambank (feature GB025), the ordering of nouns and adjectives are represented categorically, with languages generally split into three categories: Adjective-Noun, Noun-Adjective, or No dominant order. Instead of trying to fit a given language into one of these discrete categories, we extract the proportions of Adjective-Noun and Noun-Adjective instances in that language's Universal Dependencies (UD) treebank (Nivre et al., 2020).

To do this, we iterate through all of the sentences in the given dataset, counting the number of times

for all $d \in$ UD Datasets **do**
  $na \leftarrow 0$ ▷ na is the Noun-Adj count
  $an \leftarrow 0$ ▷ an is the Adj-Noun count
  for all sentence $s \in d$ **do**
    $na \leftarrow na+$ **count** Noun-Adj in s
    $an \leftarrow an+$ **count** Adj-Noun in s
  end for
  $na\_proportion \leftarrow \frac{na}{na+an}$
end for

Figure 2: Pseudocode depicting our process of collecting data for one linguistic feature.

adjectives occur before the noun they modify, as well as the number of times they occur after the noun they modify. Two examples can be seen in Table 1, where the phrase *Mon cher ami* adds one to the Adjective-Noun count, and the phrase *Mon appartement ancien* adds one to the Noun-Adjective count. We then use those counts to calculate the proportion of Adjective-Noun vs. Noun-Adjective instances that occur in the dataset.

We repeat this process for every dataset in UD that includes the necessary Noun and Adjective part of speech annotations. This algorithm is described in pseudocode in Figure 2. Because some languages have multiple datasets in UD, these languages have multiple Adjective-Noun and Noun-Adjective proportion datapoints. In the case of our seed dataset, we were able to extract information from 132 different UD datasets, within which there are 91 unique languages.

For this seed dataset, we extract data for five features:

1. Ordering of adjectives and their nouns
2. Ordering of numerals and their nouns
3. Ordering of subjects and verbs
4. Ordering of objects and verbs
5. Ordering of objects and subjects

Each feature required manual adjustments of the dataset creation code in order to extract the necessary part of speech information from the annotated UD data. These changes are small overall, generally requiring only an adjustment of the UD tags being matched. The tags we used can be found in Table 4 of Appendix A.

3.2 Value Distributions

As Figure 1 demonstrates, each feature's data creates a different distribution across the range of possible proportions. Using these raw proportions allows us to observe linguistic differences between languages that would previously be collapsed into the same category. This is made especially clear by the visualization of WALS data (black) in Figure 1, which is a much more limited distribution than its Noun Adjective counterpart in yellow.

4 Proposed Task and Model Comparison

Because categorical typological datasets are a core part of many existing typology-related NLP tasks, these tasks also suffer from many of the problems that the underlying datasets do. Examples of these tasks include typological feature prediction (Malaviya et al., 2017; Bjerva et al., 2020; Bjerva, 2024), low-resource language vocabulary prediction (Rani et al., 2023), and language identification from speech (Salesky et al., 2021). It is for this reason that we introduce, along with the seed dataset, a new task predicting these novel continuous typological features. Unlike previous typological prediction tasks, the one we present here is regression-based.

4.1 Methodological Comparison

Most typological feature prediction (TFP) approaches use logistic regression (e.g. Malaviya et al., 2017; Bjerva and Augenstein, 2018a,b; Östling and Kurfalı, 2023), as they are modelling categorical outcome variables. However, we argue that linear regression is a more suitable method for TFP, since a more appropriate representation of typology is continuous (Levshina et al., 2023). To quantify the differences between these approaches, we compare prediction results based on pretrained language vectors from Östling and Tiedemann (2017) and Malaviya et al. (2017).

As a baseline, we train logistic regression models on a discretized version of the word order features from our dataset. We have rounded each proportion to 0 or 1 (with all numbers 0.5 and above going to 1), to simulate a still-categorical version of the data, while ensuring comparability with the linear regression data. In this case, we use the following:

$$\boldsymbol{Y} = \frac{1}{1 + e^{(-\boldsymbol{\beta}\boldsymbol{X} - \boldsymbol{\beta}_0)}}$$

where $\boldsymbol{X}$ is a matrix made up of pretrained language vectors, $\boldsymbol{Y}$ is a vector made up of the in-

	Östling Linear Regr.	Östling Logistic Regr.	Malaviya Linear Regr.	Malaviya Logistic Regr.
Noun-adjective	0.146	0.261	0.141	0.378
Noun-numeral	0.140	0.132	0.129	0.399
Subject-verb	0.0781	0.306	0.101	0.156
Object-verb	0.169	0.237	0.0757	0.122
Object-subject	0.0127	–	0.0349	0.00940

Table 2: Mean squared error scores for linear regression and logistic regression models for each feature, using language vectors from Östling and Tiedemann (2017) and Malaviya et al. (2017). Better scores are closer to 0.

	Östling Linear Regr.	Östling Logistic Reg.	Malaviya Linear Regr.	Malaviya Logistic Regr.
Noun-adjective	-0.0423	-1.41	0.0810	-0.780
Noun-numeral	0.246	-3.15	-14.0	-2.45
Subject-verb	-0.233	-1.21	-0.627	-0.776
Object-verb	-0.137	-3.12	0.00891	-0.486
Object-subject	-0.299	–	-0.277	-1.84

Table 3: r^2 scores for linear regression and logistic regression models for each feature, using language vectors from Östling and Tiedemann (2017) and Malaviya et al. (2017). Better scores are closer to 1.

put language vectors' corresponding typological feature values, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ are the learned parameters. We employ the Scikit-learn (Pedregosa et al., 2011; Buitinck et al., 2013) implementation, which aims to find the optimal values of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ by minimizing the log likelihood of the data.

As an alternative approach, we train linear regression models on the language representations and use our gradient word order typology labels. For the modelling, we use:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{X}$ is again a matrix made up of pretrained language vectors, $\boldsymbol{Y}$ is again a vector made up of the input language vectors' corresponding typological feature values, $\boldsymbol{\beta}$ is the vector of learned regression coefficients, and ε is the bias vector. We use the Scikit-learn (Pedregosa et al., 2011; Buitinck et al., 2013) implementation of linear regression to train the model, which does so by minimizing the residual sum of squares between the real feature values and the predicted feature values.

For all models, both linear and logistic, we trained on a subset of the available languages, and display results, measured both in mean squared error and r^2 score, calculated on a held-out test set. Because we employed pretrained language vectors as part of the training process, we were only able to train and evaluate each feature model on the set of languages that had both a pretrained language vector, and a value in our dataset for that feature. Unfortunately, this meant that our training set for each model had only around 40 datapoints, while our held-out evaluation set had only around 10 (with some slight variation depending on the feature and the language vector source). In cases where these languages had multiple available treebanks, we randomly selected one treebank to use, to avoid training on the same input vectors with potentially different expected output feature values. We selected one treebank randomly instead of combining them into one set per language so as to not arbitrarily combine data from potentially vastly different domains. Detailed results are displayed in Tables 2 and 3.

4.2 Results and Discussion

Given that the data at hand is continuous, and that linear regression models predict categorical values while logistic regression models predict binary values, we expected the linear regression models to outperform the logistic regression models on this task. Indeed, the linear regression models perform better on average than the logistic regression models, when evaluated using mean squared error and r^2 score. While not always the case, this is most often true as well on the individual feature level. While improvements to the modelling can be implemented, these baselines serve as an initial exploration of how to approach the novel task of regression-based typology prediction.

An important note from our statistical results is that the differences we observe between the data driven distributions and typological databases (Fig. 1) clearly show the limitations of established databases in terms of language descriptiveness

on a fine-grained scale. This discrepancy may to some extent explain the difficulty observed in empirical NLP experiments, when trying to integrate coarse-level WALS features in various NLP pipelines (Ponti et al., 2019). The introduction of this regression-based typology prediction task may prove useful for incorporation of typological features in NLP modelling - for instance by incorporation as an auxiliary task.

While data-driven typology enables more fine-grained language description, it should be noted that the source of a treebank can have a considerable effect on the estimate (Levshina et al., 2023). Baylor et al. (2023) show that linguistic variation, for instance stemming from domain, can affect word order values. Therefore, direct comparison between languages should ideally be based on parallel data.

5 Conclusion

Information from the field of linguistic typology has the potential to benefit the field of NLP. Unfortunately, the data from existing typological databases has been unreliable, largely due to their reliance on categorical features and those features' inability to represent the variability found in natural language. In this paper, we attempt to address this problem by introducing a new continuous-valued seed dataset, and argue that it is indeed better able to reflect the nuance of natural language when it comes to word order. In addition, we provide our dataset creation methodology that can be easily adapted in the future to generate data for a wider array of languages and features. Finally, we present a novel regression task based on predicting the feature values of this new dataset.

Limitations

The main limitation of our paper stems from the small size of our dataset, both in terms of number of features, and in terms of languages covered. As is always possible, our subset of features and languages could be misrepresentative of the larger existing features and languages, thus keeping our analyses from generalizing. The small size of our dataset only makes this more probable.

A secondary limitation of this work primarily applies to our dataset creation method. As it currently stands, the method only works with annotated linguistic data, vastly cutting down on the amount of available useful language data.

Ethics Statement

As this paper relies on existing linguistic data sources from which to generate datasets, no human data was collected.

We do not foresee this work directly creating any substantial ethical issues, but we do note that language communities can be significantly impacted, both positively and negatively, by language technologies. Given that this research has the potential to aid in the further development of language technologies, we want to highlight the importance of community-led development, including ceasing development of technologies for certain languages based on community request.

Acknowledgements

This work was supported by the Carlsberg Foundation under the *Semper Ardens: Accelerate* programme (CF21-0454). EB was further supported by the McGill University Graduate Mobility Award to travel to AAU to carry out this work.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1163–1169, Singapore. Association for Computational Linguistics.

Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Johannes Bjerva. 2024. The Role of Typological Feature Prediction in NLP and Linguistics. *Computational Linguistics*, pages 1–14.

Johannes Bjerva and Isabelle Augenstein. 2018a. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018b. Tracking Typological Traits of Uralic Languages in Distributed Language Representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

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.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Bernard Comrie. 1988. Linguistic typology. *Annual Review of Anthropology*, 17:145–159.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Matthew S. Dryer. 2013. Order of subject and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

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

Joseph H. Greenberg. 1966. *Universals of Language*. The MIT Press.

Larry M. Hyman. 2008. Universals in phonology.

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*, 0(0).

Björn Lindblom and Ian Maddieson. 1988. Phonetic universals in consonant systems. *Language, speech and mind*, 6278.

Ian Maddieson. 2013. Vowel quality inventories (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Pado. 2022. Word-order typology in multilingual BERT: A case study in subordinate-clause detection. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 11–21, Seattle, Washington. Association for Computational Linguistics.

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 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is 'Typological Diversity' in NLP?

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.

Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns, and John P. McCrae. 2023. Findings of the SIGTYP 2023 shared task on cognate and derivative detection for low-resourced languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 126–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, and Monojit Choudhury. 2014. Automatic discovery of adposition typology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1037–1046, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Elizabeth Salesky, Badr M. Abdullah, Sabrina Mielke, Elena Klyachko, Oleg Serikov, Edoardo Maria Ponti, Ritesh Kumar, Ryan Cotterell, and Ekaterina Vylomova. 2021. SIGTYP 2021 shared task: Robust spoken language identification. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 122–129, Online. 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, Tânia 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 the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

A Tags for algorithm beyond Adjective-Noun order

POS	UD upos value	UD deprels value
Noun	NOUN	–
Adjective	ADJ	amod
Numeral	NUM	nummod
Subject	–	nsubj
Object	–	obj
Verb	VERB	–

Table 4: Tags used to extract the necessary parts of speech from the Universal Dependencies treebank (Nivre et al., 2020). Dashes indicate that that value did not need to be specified.