Squib

The Role of Typological Feature Prediction in NLP and Linguistics

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

Computational typology has gained traction in the field of Natural Language Processing (NLP) in recent years, as evidenced by the increasing number of papers on the topic and the establishment of a Special Interest Group on the topic (SIGTYP), including the organization of successful workshops and shared tasks. A considerable amount of work in this sub-field is concerned with prediction of typological features, for example, for databases such as the World Atlas of Language Structures (WALS) or Grambank. Prediction is argued to be useful either because (1) it allows for obtaining feature values for relatively undocumented languages, alleviating the sparseness in WALS, in turn argued to be useful for both NLP and linguistics; and (2) it allows us to probe models to see whether or not these typological features are encapsulated in, for example, language representations. In this article, we present a critical stance concerning prediction of typological features, investigating to what extent this line of research is aligned with purported needs—both from the perspective of NLP practitioners, and perhaps more importantly, from the perspective of linguists specialized in typology and language documentation. We provide evidence that this line of research in its current state suffers from a lack of interdisciplinary alignment. Based on an extensive survey of the linguistic typology community, we present concrete recommendations for future research in order to improve this alignment between linguists and NLP researchers, beyond the scope of typological feature prediction.

1. Introduction

Over the course of the past two centuries, linguistic typologists have studied languages with respect to their structural and functional properties, thereby implicitly classifying languages as being more or less similar to one another by virtue of such properties (Comrie 1988; Haspelmath et al. 2001; Velupillai 2012). Typology has a long history (Herder 1772; Gabelentz 1891; Greenberg 1960, 1974; Dahl 1985; Comrie 1989; Croft 2003), and recently computational approaches have gained substantial popularity (Wichmann and Saunders 2007; Dunn et al. 2011; Wälchli 2014; Östling 2015; Cotterell

Submission received: 18 August 2023; accepted for publication: 7 October 2023.

https://doi.org/10.1162/coli_a_00498

and Eisner 2017; Asgari and Schütze 2017; Malaviya, Neubig, and Littell 2017; Bjerva and Augenstein 2018b; Levshina 2019; Bjerva et al. 2020; Oncevay, Haddow, and Birch 2020; Östling and Kurfalı 2023; Baylor, Ploeger, and Bjerva 2023). One part of traditional typological research deals with manually extracting features of languages from existing descriptions, for instance, ending up in databases such as the World Atlas of Language Structures (WALS, Dryer and Haspelmath 2013), URIEL (Littell et al. 2017), AUTOTYP (Bickel et al. 2023), PHOIBLE (Moran and McCloy 2019), and most recently Grambank (Skirgård et al. 2023). A recent development that can be seen as complementary to this is the process of learning distributed language representations in the form of dense real-valued vectors, often referred to as **language embeddings** (Tsvetkov et al. 2016; Östling and Tiedemann 2017; Malaviya, Neubig, and Littell 2017; Jin and Xiong 2022; Bjerva et al. 2019c; Harvill, Girju, and Hasegawa-Johnson 2022; Chen, Biswas, and Bjerva 2023).

In this article, we focus on the task of typological feature prediction, as introduced by work such as Teh, Daumé III, and Roy (2009) and Daumé III and Campbell (2007), and featured in the SIGTYP 2020 Shared Task (Bjerva et al. 2020). Once a relatively niche topic in the NLP community, studying typological features has recently risen in popularity and importance for a number of reasons. The field has seen considerable advances in cross-lingual transfer learning, whereby stable cross-lingual representations can be learned on massive amounts of data in an unsupervised way, be it for words (Ammar et al. 2016; Wada, Iwata, and Matsumoto 2019) or sentences (Artetxe and Schwenk 2019; Devlin et al. 2019; Conneau and Lample 2019; Conneau et al. 2020; Tiyajamorn et al. 2021; Ouyang et al. 2021). This naturally raises the question of what these representations encode, and some have turned to typology for potential answers (Choenni and Shutova 2020; Zhao et al. 2021; Stanczak et al. 2022). In a similar vein, research has shown that these learned representations can be fine-tuned for supervised tasks, then applied to new languages in a few- or even zero-shot fashion with surprisingly high performance. This has raised the question of what causes this performance, and to what degree typological similarities are exploited by such models (Bjerva and Augenstein 2018a; Nooralahzadeh et al. 2020; Zhao et al. 2021; Östling and Kurfalı 2023). In addition to using typology for diagnostic purposes, prior work has also found that typology can, to some extent, guide cross-lingual sharing (de Lhoneux et al. 2018). Finally, the relationship between typological resources such as WALS (Dryer and Haspelmath 2013) and language representations has been studied, which has shown that knowledge base population methods can be used to complete typological resources (Malaviya, Neubig, and Littell 2017; Murawaki 2017; Bjerva and Augenstein 2018a; Bjerva et al. 2019c), and that typological implications can be discovered automatically (Daumé III and Campbell 2007; Bjerva et al. 2019b). Experiments in using typological features for NLP typically find sporadic and limited benefits (O'Horan et al. 2016; Ponti et al. 2019; Oncevay, Haddow, and Birch 2020).

While many such applications are well-motivated, the precise purpose of *predicting* typological features remains unclear. In this article, we investigate this question, provide an overview of arguments used in the NLP literature, and assess these arguments critically. In order to address this question, we first provide an overview of past work and current usage areas of typology and typological feature prediction in NLP. We next turn to linguistics, and present results of a survey and in-depth interviews of experts in typology, experts in language documentation, and other linguists, in order to map out the usefulness of our current work. Finally, we give recommendations on future research directions based on our findings, in an attempt to improve alignment between work in computational linguistics focused on typological feature prediction, and what may actually be of use to field linguists and typologists.

2. Related Work

We present a brief overview of typological feature prediction and its uses in NLP here, and refer the reader to Ponti et al. (2019) for a more thorough overview focusing on empirical usefulness of typological feature prediction. In the context of NLP, typological feature prediction is commonly done in the context of existing databases (e.g., WALS, Dryer and Haspelmath 2013), or more recently in Grambank (Skirgård et al. 2023). Methodologically speaking, features are typically either used or predicted in the context of other features and other languages (Teh, Daumé III, and Roy 2009; Daumé III and Campbell 2007; Naseem, Barzilay, and Globerson 2012; Täckström, McDonald, and Nivre 2013; Berzak, Reichart, and Katz 2014; Malaviya, Gormley, and Neubig 2018; Bjerva et al. 2019c, 2019a, 2020, 2019b; Vastl, Zeman, and Rosa 2020; Jäger 2020; Choudhary 2020; Gutkin and Sproat 2020; Kumar et al. 2020). That is to say, given a language $l \in L$, where L is the set of all languages contained in a specific database, and the features of that language F_l, the setup is typically to attempt to predict some subset of features $f \subset F_l$, based on the remaining features $F_l \setminus f$. This language may be (partially) held out during training, such that a typological feature prediction model is fine-tuned on $L \setminus l$, before being evaluated on language l. Variations of this setup exist, with attempts to control for language relatedness in training/test sets, using genealogical, areal, or structural similarities (Bjerva et al. 2020; Östling and Kurfalı 2023). In general, the degree to which areal and genealogical factors are controlled for in typological feature prediction is quite limited. Typically, previous work attempts to hold out languages during training in a given radius of, for example, 1,000 km (Jaeger et al. 2011; Cysouw 2013; Bjerva et al. 2020), or attempt to use family and branch information to avoid overestimation of prediction power. Related work either follows this type of approach (Östling and Kurfalı 2023), or omits controls altogether. While not the core of this article, a general recommendation is that future work take this type of factor into account—for example, by using linguistically motivated filtering approaches based on macroareas (for example) (Dryer 1989, 1992; Hammarström and Donohue 2014; Miestamo, Bakker, and Arppe 2016), the somewhat more fine-grained AUTOTYP areas (Nichols and Bickel 2009) which include historical, genetic, archaeological and anthropological factors, sociolinguistic environments (Sinnemäki and Di Garbo 2018), or using information regarding shared borders between languages (Cysouw, Dediu, and Moran 2012; Dryer 2018).

3. Why Do NLP Practitioners Predict Typological Features?

The adoption of the task of typological feature prediction in NLP stems from three core arguments in the literature: (1) sparsity, (2) continuity, and (3) utility for NLP. Although these arguments are frequently made, we here argue that they are largely unsubstantiated.

3.1 Sparsity: "Typological Databases Are Sparse and Incomplete"

Many typological databases indeed contain gaps for feature-language combinations. This certainly is the case with, for example, WALS and URIEL, where many combinations are absent. Many gaps exist for good reasons—the WALS feature NASAL VOWELS IN WEST AFRICA is, for obvious reasons, absent for languages outside of West Africa. Some databases, such as Phoible and Grambank, generally do not suffer from this particular issue (Skirgård et al. 2023). There is a general argument echoed by, for

example, Daumé III and Campbell (2007), Berzak, Reichart, and Katz (2014), Buis and Hulden (2019), and Bjerva et al. (2019a), stating that completing these databases would be useful for typologists, highlighting that it is a difficult task to solve. Furthermore, it is argued that inaccurate information in databases can be detected and fixed automatically.

While imputing missing data can be useful for downstream tasks in, for example, computational historical linguistics, we here consider whether such predictions constitute a contribution to typological knowledge. Generally speaking, the predictions made by such systems for undocumented features in WALS are rather well-known. This stems from the core issue with such methods, namely, that they are first and foremost based on correlations, be it between typological features (e.g., affixation correlates with basic word order), or between similar languages (e.g., most Germanic languages are SVO). As models are typically good at picking up on such correlations, one of the findings in the SIGTYP2020 shared task was that practically all system submissions are able to correctly predict *easy* features. In the case of more difficult features (e.g., rare or atypical combinations), the best models only attained an accuracy of roughly 65% (Bjerva et al. 2020). Hence, in the cases where a language is typologically *interesting* (e.g., where an uncommon combination of typological features occurs), current state-of-the-art models do not fare well.

3.2 Continuity: "NLP Can Facilitate a Continuous Scale View on Typology"

An argument with support in the linguistic literature deals with the fact that, for example, word-order typology arguably lies on a continuum, rather than discrete categorization (Levshina et al. 2023). For instance, French allows for both Noun-Adjective and Adjective-Noun ordering, depending on various constraints (Laenzlinger 2005). An empirical investigation of word-order typology in the Universal Dependencies dataset provides a detailed cross-lingual perspective on the matter. Following Baylor, Ploeger, and Bjerva (2023), we use dependency links to calculate the proportion of, for example, Noun-Adjective vs. Adjective-Noun ordering examples across 100 languages. Contrasting this with categorical features, as represented in WALS, highlights the fact that this type of representation is a poor match with the feature distributions seen across corpora (Figure 1). Clearly, basic approaches to computational linguistics can help paint a descriptive picture of language data in this manner. However, would an output from a black-box NLP model, saying that a language is "40% Noun-Adjective," be useful, or is a more descriptive and transparent method, as described, required?

3.3 Utility: "Prediction of Typological Features Can Be Useful for NLP"

Finally, it is commonly argued that typological features can aid performance in multilingual NLP models, for example, serving as a guide in cross-lingual transfer (Lent et al. 2023). Indeed, limited benefits can be found in various experimental setups across common NLP tasks and languages with annotated features (Naseem, Barzilay, and Globerson 2012; Täckström, McDonald, and Nivre 2013; de Lhoneux et al. 2018), and previous work has shown that typological information is learned as a by-product of training (Bjerva and Augenstein 2021). As Figure 1 hints, it may also be that the culprit is the inherent mismatch between typological database information and data-driven gradient typology (Baylor, Ploeger, and Bjerva 2023, 2024). Considering *predicted* typological features, Üstün et al. (2022) find benefits in zero-shot settings for parsing. However, work considering typological similarities when finding appropriate language pairings in cross-lingual transfer often finds combinations which are not easily explained by

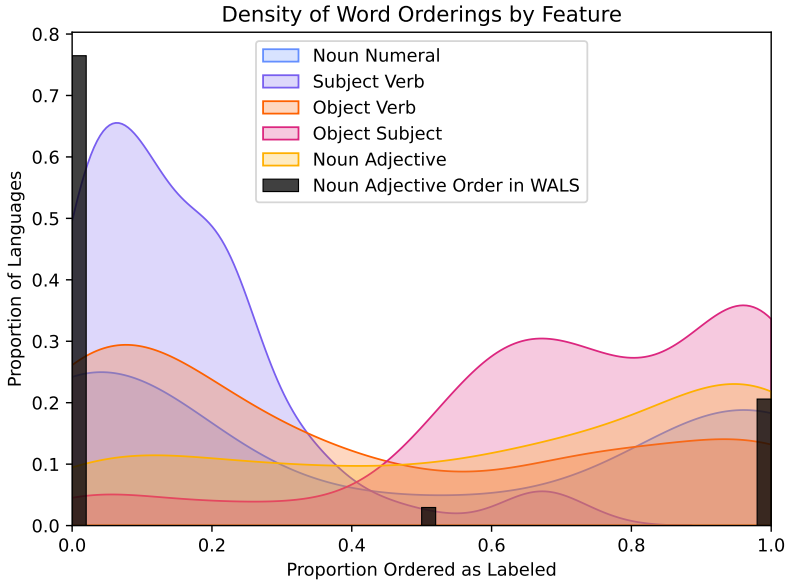

Figure 1
Distribution of word-order features across a sample of 100 languages in the Universal Dependencies dataset (De Marneffe et al. 2021). Black bars represent categorical feature values as represented in WALS, and lines show the distribution of expressions in UD. The proportion on the *x*-axis follows the ordering from each label—e.g., 0.0 "Noun Adjective" means entirely "Noun Adjective," while 1.0 means entirely "Adjective Noun."

typology, likely due to artifacts in training or evaluation setups (Dolicki and Spanakis 2021; de Vries, Wieling, and Nissim 2022). In this vein, Srinivasan et al. (2021) contend that low performance for Yoruba may be due to its vigesimal number system, whereas m-BERT is primarily trained on languages using the decimal system—it is difficult to substantiate that this is much more than a spurious correlation. In sum, although there is debate on the subject of utility, we argue that this argument is likely the only valid reason for predicting typological features, as it stands today.

4. Do Linguists Want Typological Feature Prediction?

Having established the common arguments for prediction of typological features used by the NLP community, we now turn to the linguistic community to investigate these claims. Do linguists agree that typological feature prediction constitutes a contribution to the field, solving an inherently difficult problem? In terms of difficulty, it appears at first glance that this might be the case. For instance, Dryer (2007) points out that "it may be difficult to distinguish pronouns from nouns except on a semantic basis", and Curnow (2000) argues that it is difficult to distinguish between *inflectional* and *zero* copulas in languages without verbal morphology. However, Haspelmath (2021) outlines an important distinction in this area. It is not that drawing distinctions between typological categorization is difficult, but rather that there is an underlying data issue making it difficult to draw sound conclusions based on a sufficient sample.

The literature does not have much to say about the usefulness of the task, however. Based on this, we have developed a questionnaire to investigate whether this line of research is useful to linguists, and if not, what needs to be changed so as to provide utility. The design of the questionnaire focused on the core research question of this article in mind, aiming to tease apart whether what NLP is currently doing is useful and, if not, what *might* be useful for future work. Specifically, the findings in this section are based on a survey and in-depth interviews with experts in linguistic typology, language documentation, and general linguistics. The survey was disseminated among experts on the *Lingtyp* mailing list, and directly to several linguistics departments worldwide, including follow-up interviews with linguists at various career stages. The respondents were initially informed of the survey's scope:

> In recent years, the field(s) of Natural Language Processing (NLP) and Computational Linguistics (CL) have started paying increased amounts of attention to linguistic typology. Among other things, NLP/CL researchers have developed systems for automatic inference of typological features. Typically, NLP/CL researchers working with typological features claim that this research direction has potential relevance to linguistics. However, it is not established that this line of research has any relevance to linguists at all. In this survey, we aim to bridge the gap between NLP/CL and linguistics researchers. Initially we want to create an overview of how linguists perceive such NLP/CL efforts, to what extent they are useful, or may be useful to the field in the future. The end goal is to improve alignment of research efforts of NLP/CL researchers with an interest in, e.g., typology, with the actual needs of the linguistic community.

4.1 Survey Respondents

The survey attracted a total of 34 responses, across career stages, with representation from 20 countries, on 3 continents. Out of the surveyed population, 80% identify as being linguistic typology researchers, 70% as working with language documentation, and 60% as working with general linguistics. Eighty percent of respondents are at a postdoctoral stage or later, with the remaining 20% being graduate students, bachelor's students, or other.

4.2 Quantitative Responses

Following this prompt above, an initial survey was carried out in which respondents provided answers on a 5-point Likert scale, with descriptors at each end point (*Not at all useful* – *Highly useful*). The following questions were provided, with summaries of responses in Table 1:

1. Is automated prediction of features based on other known features useful?
2. Is prediction based on descriptions of language, e.g., grammars, useful?
3. Is prediction from textual input in a language, e.g., collected and transcribed samples, useful?
4. Is prediction from sound input in a language, e.g., recorded speech, useful?
5. How important is **explainability** in the utility of the models?

Table 1
Summary of responses to questions on typological feature prediction (TFP) by linguistic typologists. Responses were given on an ordinal scale from 1 (lowest) to 5 (highest).

Question	Mode	Median	Distribution	Summary
1. TFP from features	2	2		Not useful
2. TFP from grammars	4	4		Partially useful
3. TFP from transcriptions	3	4		Partially useful
4. TFP from speech	4	4		Partially useful
5. Explainability	5	5		Very important

The general trend in the survey responses is provided in Table 1, which is generally symptomatic of a lack of alignment between NLP practitioners and linguists. All approaches to TFP are found to be not useful, or moderately useful. Explainability is highlighted as a key feature for the success of any TFP tool.

4.3 Qualitative Responses

In addition to these questions, qualitative responses were gathered in part from a free-text input in the questionnaire, in addition to a series of semi-structured interviews with experts in the community. Generally speaking, the qualitative responses gathered tell a story of skepticism. NLP practitioners are viewed as neglecting the efforts of language documentation, without much understanding of basic documentation workflow, highlighting a need for us as a community to get a grasp of this before commenting on it. While many responses indicate that NLP/CL researchers offer valuable feedback to the linguistic community, for example, in facilitating access to automatic speech recognition for corpora creation, the specific aspect of typological feature prediction generally does not seem to be particularly valued. Indeed, the surveyed population also point out the well-established aspect of the problem of categorical values in typological databases, for example, stating that language descriptions are better formulated as "language X has category Y, but …", or "morphosyntactic pattern X is attested in language Y, but …"

5. The Future of Typological Feature Prediction

Based on the survey, we here propose three concrete directions for future work.

5.1 Make Predictions Explainable

Explainability is a crucial factor in typological feature prediction, particularly if the goal is for predicted features to be useful for typologists. Both quantitative and qualitative survey responses indicate that specific attributions of feature predictions are needed, e.g., via indication of specific examples in grammars or transcriptions that verify any claims made. This echoes findings in other work on acceptance of artificial intelligence (AI), specifically in that explainability is the key to AI acceptance (Shin 2021).

Concretely, methodologies based on saliency metrics or contrastive learning between typologically distinct languages may be useful avenues to explore in future NLP research incorporating TFP.

5.2 Communicate with Domain Experts

The issue outlined in this article is one of misalignment between communities, essentially instantiation of a long-standing issue in NLP, commonly referred to as a pendulum oscillating between a linguistic focus, and an engineering focus (Church and Liberman 2021). Typological feature prediction has, perhaps, seemed like a task with clear utility to a specific community, due to its inherent "difficulty." However, as outlined in this article, and as argued by, for example, Haspelmath (2021), the difficulty is not in categorization of languages into specific feature buckets, but rather one of data scarcity. Concretely, we suggest that future work that aims to have relevance to the linguistic community is spurred by *communication with domain experts*. Linguistics offers rigorous frameworks for understanding the intricate properties and structures inherent in human language. This theoretical foundation has found its way into many areas of NLP, and is lacking in others. Conversely, empirical findings from NLP can highlight potential research avenues within linguistics. A structured communication channel between the two domains can alleviate introduction of theoretical findings from linguistics in computational models, and empirical results from NLP can be contextualized within linguistic theories. With improved alignment, a deeper and more comprehensive understanding of both the structure and function of language can be achieved, fostering novel scientific insights.

5.3 Base Predictions on Real Data

Basing predictions on structured data, such as from existing features, is not deemed particularly insightful by the community. As highlighted, correlative predictions based on other features are typically either well-known and carry little novel value for a linguist, or are based on spurious correlations and entirely nonsensical. Concretely, development of a typological feature prediction system that uses text or meta-text as input might have significant value to the community, if paired with explainability. For instance, correctly analyzing a language as being suffixing in its inflectional morphology, while pointing to concrete examples of such suffixing, is an example with potential value.

6. Conclusions

For years, the NLP and CL communities have touted the task of typological feature prediction as one fulfilling a specific need in the linguistic community. The results outlined in this article largely refute this claim. While any further claims to this effect should be revisited, we further recommend that other claims in the NLP community are sanity checked with regards to the group they supposedly help. This is clearly the case in interactions with linguists, but also echoes the sentiment of Bird (2021) in that, for example, some communities simply do not have an expressed need for specific language technologies. In short, future work in NLP making claims of community interaction and benefiting marginalized groups ought to invest the effort needed to verify these claims, before they become widespread and accepted without any interdisciplinary grounding.

Acknowledgments

This work was immensely improved thanks to input from discussions with Esther Ploeger, Emi Baylor, Marcell Fekete, Yiyi Chen, Heather Lent, Carl Börstell, Johan Sjons, and Bruno Olsson. We further thank the LINGTYP community for participating in the survey in this work, as well as the broad range of anonymous reviewers for their feedback on the initial version of the article. This work was supported by a *Semper Ardens: Accelerate* research grant (CF21-0454) from the Carlsberg Foundation.

References

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. https://doi.org/10.1162/tacl_a_00288

Asgari, Ehsaneddin and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of EMNLP*, pages 113–124. https://doi.org/10.18653/v1/D17-1011

Baylor, Emi, 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*. https://doi.org/10.18653/v1/2023.findings-emnlp.82

Baylor, Emi, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from universal dependencies. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. March 2024.

Berzak, Yevgeni, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29. https://doi.org/10.3115/v1/W14-1603

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zùñiga, and John B. Lowe. 2023. The AUTOTYP database (v1.1.1). https://doi.org/10.5281/zenodo.7976754

Bird, Steven. 2021. EMNLP keynote: LT4All!? Rethinking the agenda. *The 2021 Conference on Empirical Methods in Natural Language Processing*.

Bjerva, Johannes 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. https://doi.org/10.18653/v1/N18-1083

Bjerva, Johannes 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. https://doi.org/10.18653/v1/W18-0207

Bjerva, Johannes 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 (EACL)*. https://doi.org/10.18653/v1/2021.eacl-main.38

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540. https://doi.org/10.18653/v1/N19-1156

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019b. Uncovering probabilistic implications in typological knowledge bases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930. https://doi.org/10.18653/v1/P19-1382

Bjerva, Johannes, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019c. What do language representations really represent? *Computational Linguistics*, 45(2):381–389. https://doi.org/10.1162/coli_a_00351

Bjerva, Johannes, 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. https://doi.org/10.18653/v1/2020.sigtyp-1.1

Buis, Annebeth and Mans Hulden. 2019. Typological feature prediction with matrix completion. In *Proceedings of TyPNLP: The First Workshop on Typology for Polyglot NLP*, pages 13–15.

Chen, Yiyi, Russa Biswas, and Johannes Bjerva. 2023. Colex2Lang: Language embeddings from semantic typology. In *The 24rd Nordic Conference on Computational Linguistics (NoDaLiDa)*.

Choenni, Rochelle and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. *CoRR*, abs/2009.12862.

Choudhary, Chinmay. 2020. NUIG: Multitasking self-attention based approach to SigTyp 2020 shared task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. https://doi.org/10.18653/v1/2020.sigtyp-1.6

Church, Kenneth and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:625341. https://doi.org/10.3389/frai.2021.625341, PubMed: 33954287

Comrie, Bernard. 1988. Linguistic typology. *Annual Review of Anthropology*, 17:145–159. https://doi.org/10.1146/annurev.an.17.100188.001045

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS*, pages 7057–7067.

Cotterell, Ryan and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of ACL*, pages 1182–1192. https://doi.org/10.18653/v1/P17-1109

Croft, William. 2003. *Typology and Universals*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840579

Curnow, Timothy Jowan. 2000. Towards a cross-linguistic typology of copula constructions. In *Proceedings of the 1999 Conference of the Australian Linguistic society*, volume 11, pages 203–300.

Cysouw, Michael. 2013. Disentangling geography from genealogy. In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*. de Gruyter. https://doi.org/10.1515/9783110312027.21

Cysouw, Michael, Dan Dediu, and Steven Moran. 2012. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa." *Science (New York, N.Y.)*, 335:657; author reply 657. https://doi.org/10.1126/science.1208841, PubMed: 22323802

Dahl, Östen. 1985. *Tense and Aspect Systems*. Basil Blackwell Ltd., New York.

Daumé III, Hal and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72.

de Lhoneux, Miryam, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997. https://doi.org/10.18653/v1/D18-1543

De Marneffe, Marie Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

de Vries, Wietse, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685. https://doi.org/10.18653/v1/2022.acl-long.529

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Dolicki, Błażej and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *arXiv preprint arXiv:2105.05975*.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language,"* 13(2):257–292. https://doi.org/10.1075/sl.13.2.03dry

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language*, 68(1):81–138. https://doi.org/10.1353/lan.1992.0028

Dryer, Matthew S. 2007. Noun phrase structure. In *Language Typology and Syntactic Description*, volume 2, pages 151–205. https://doi.org/10.1017/CBO9780511619434.003

Dryer, Matthew S. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language*, 94(4):798–833. https://doi.org/10.1353/lan.2018.0054

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

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82. https://doi.org/10.1038/nature09923, PubMed: 21490599

Gabelentz, Georg von der. 1891. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig.

Greenberg, Joseph. 1974. *Language Typology: A Historical and Analytic Overview*, volume 184. Walter de Gruyter. https://doi.org/10.1515/9783110886436

Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3):178–194. https://doi.org/10.1086/464575

Gutkin, Alexander and Richard Sproat. 2020. NEMO: Frequentist inference approach to constrained linguistic typology feature prediction in SIGTYP 2020 shared task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. https://doi.org/10.18653/v1/2020.sigtyp-1.3

Hammarström, Harald and Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change*, 4(1):167–187.

Harvill, John, Roxana Girju, and Mark Hasegawa-Johnson. 2022. Syn2Vec: Synset colexification graphs for lexical semantic similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5270. https://doi.org/10.18653/v1/2022.naacl-main.386

Haspelmath, Martin. 2021. Typological classification is never "difficult" — the difficulties lie elsewhere. Available online at https://dlc.hypotheses.org/2528

Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, editors. 2001. *Language Typology and Language Universals: An International Handbook*. volume 20. Walter de Gruyter. https://doi.org/10.1515/9783110171549.2.12.1380

Herder, J. 1772. *Abhandlung über den Ursprung der Sprache*. Berlin: Christian Friedrich Voß.

Jaeger, T. Florian, Peter Graff, William Croft, and Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, 15:281–320. https://doi.org/10.1515/lity.2011.021

Jäger, Gerhard. 2020. Imputing typological values via phylogenetic inference. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. https://doi.org/10.18653/v1/2020.sigtyp-1.5

Jin, Renren and Deyi Xiong. 2022. Informative language representation learning for massively multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5158–5174.

Kumar, Ritesh, Deepak Alok, Akanksha Bansal, Bornini Lahiri, and Atul Kr. Ojha. 2020. KMI-Panlingua-IITKGP at SIGTYP2020: Exploring rules and hybrid systems for automatic prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. https://doi.org/10.18653/v1/2020.sigtyp-1.2

Laenzlinger, Christopher. 2005. French adjective ordering: Perspectives on DP-internal movement types. *Lingua*, 115(5):645–689. https://doi.org/10.1016/j.lingua.2003.11.003

Lent, Heather, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Hans Erik Heje, Diptesh Kanojia,

Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2023. CreoleVal: Multilingual multitask benchmarks for creoles. *arXiv preprint arXiv:2310.19567*.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572. https://doi.org/10.1515/lingty-2019-0025

Levshina, Natalia, 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*. https://doi.org/10.1515/ling-2021-0098

Littell, Patrick, 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. https://doi.org/10.18653/v1/E17-2002

Malaviya, Chaitanya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663. https://doi.org/10.18653/v1/P18-1247

Malaviya, Chaitanya, 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. https://doi.org/10.18653/v1/D17-1268

Miestamo, Matti, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296. https://doi.org/10.1515/lingty-2016-0006

Moran, Steven and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Murawaki, Yugo. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461.

Naseem, Tahira, 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: Long Papers-Volume 1*, pages 629–637.

Nichols, Johanna and Balthasar Bickel. 2009. The AUTOTYP genealogy and geography database: 2009 release. https://github.com/autotyp/autotyp-data

Nooralahzadeh, Farhad, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of EMNLP*, pages 4547–4562. https://doi.org/10.18653/v1/2020.emnlp-main.368

Oncevay, Arturo, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406. https://doi.org/10.18653/v1/2020.emnlp-main.187

Östling, Robert. 2015. Word order typology through multilingual word alignment. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 205–211. https://doi.org/10.3115/v1/P15-2034

Östling, Robert and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations. *Computational Linguistics*, pages 1–46. https://doi.org/10.1162/coli_a_00491

Östling, Robert and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *EACL*, pages 644–649. https://doi.org/10.18653/v1/E17-2102

Ouyang, Xuan, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38. https://doi.org/10.18653/v1/2021.emnlp-main.3

O'Horan, Helen, 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.

Ponti, Edoardo Maria, 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. https://doi.org/10.1162/coli_a_00357

Shin, Donghee. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sinnemäki, Kaius and Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology*, 9:1141. https://doi.org/10.3389/fpsyg.2018.01141, PubMed: 30154738

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175. https://doi.org/10.1126/sciadv.adg6175, PubMed: 37075104

Srinivasan, Anirudh, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual NLP models. *arXiv preprint arXiv:2110.08875*.

Stanczak, Karolina, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598. https://doi.org/10.18653/v1/2022.naacl-main.114

Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071.

Teh, Y. W., H. Daumé III, and D. Roy. 2009. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems 20, Proceedings of the 2007 Conference*.

Tiyajamorn, Nattapong, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774. https://doi.org/10.18653/v1/2021.emnlp-main.612

Tsvetkov, Yulia, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. *NAACL-HLT*, pages 1357–1366. https://doi.org/10.18653/v1/N16-1161

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592. https://doi.org/10.1162/coli_a_00443

Vastl, Martin, Daniel Zeman, and Rudolf Rosa. 2020. Predicting Typological Features in WALS using Language Embeddings and Conditional Probabilities: ÚFAL Submission to the SIGTYP 2020 Shared Task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. https://doi.org/10.18653/v1/2020.sigtyp-1.4

Velupillai, Viveka. 2012. *An Introduction to Linguistic Typology*. John Benjamins Publishing. https://doi.org/10.1075/z.176

Wada, Takashi, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of ACL (1)*, pages 3113–3124. https://doi.org/10.18653/v1/P19-1300

Wälchli, Bernhard. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, 28:355. https://doi.org/10.1515/9783110317558.355

Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica*, 24(2):373–404. https://doi.org/10.1075/dia.24.2.06wic

Zhao, Wei, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240. https://doi.org/10.18653/v1/2021.starsem-1.22