A Call for Consistency in Reporting Typological Diversity

Wessel Poelman[*♣] **Esther Ploeger**[*✿] **Miryam de Lhoneux**[♣] **Johannes Bjerva**[✿]
[♣]Department of Computer Science, KU Leuven, Belgium
[✿]Department of Computer Science, Aalborg University, Denmark
{wessel.poelman,miryam.delhoneux}@kuleuven.be {espl,jbjerva}@cs.aau.dk

1 Introduction

In order to draw generalizable conclusions about the performance of multilingual models across languages, it is important to evaluate on a set of languages that captures linguistic diversity. Linguistic typology is increasingly used to justify language selection, inspired by language sampling in linguistics (e.g., Rijkhoff and Bakker, 1998). In other words, more and more papers suggest generalizability by evaluating on 'typologically diverse languages' (see Figure 1). However, justifications for 'typological diversity' exhibit great variation, as there seems to be no set definition, methodology or consistent link to linguistic typology. In this work, we provide a systematic insight into how previous work in the ACL Anthology uses the term 'typological diversity'. Our two main findings are:

1. What is meant by typologically diverse language selection is not consistent.
2. The actual typological diversity of the language sets in these papers varies greatly.

We argue that, when making claims about 'typological diversity', an operationalization of this should be included. A systematic approach that quantifies this claim, also with respect to the number of languages used, would be even better.

2 Systematic Annotation of Claims

We systematically investigate which papers make claims regarding typological diversity, and which languages they actually use. First, we retrieve[1] all papers in the ACL Anthology that contain the following search string in either the title or abstract:

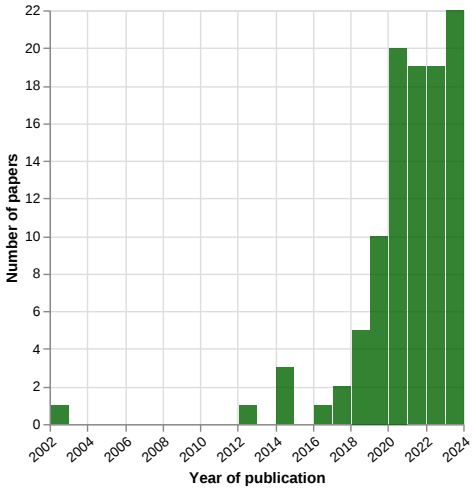

Figure 1: Number of papers in the ACL Anthology claiming a 'typologically diverse' set of languages over the years.

```
typological.+?diverse|
typological.+?diversity|
diverse.+?typological
```

Examples of this are not only *typologically diverse*, but also *typologically maximally diverse language* and *typologically and genetically diverse languages*. In total, this retrieves 140 papers, with the earliest being published in 2002, and the most recent being published in 2023. It contains papers from conferences (e.g., *ACL, EMNLP), journals (e.g., TACL, CL) and workshops (e.g., SIGTYP, SIGMORPHON).

We manually annotate whether these papers contain a claim regarding the typological diversity of their language selection. An example of such a claim is: *"we evaluate on a set of ten typologically diverse languages"* (Pimentel et al., 2020). A paper does not make a claim if it describes related work that claims to use 'a diverse typological test set', for instance. Our annotation is done separately by two annotators (the first two authors). We calculate inter-annotator agreement and retrieve a Cohen's κ of 0.64 ('substantial agreement'). After resolving the disagreements, we are left with 103 papers that

* Equal contribution.
[1]Using the `acl-anthology-py` package: https://github.com/mbollmann/acl-anthology-py. Papers retrieved on December 11, 2023.

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 75–77

contain a claim, which we use for our analysis. For every such paper, we annotate which languages are actually included in their selection. We normalize these to ISO-639-3 codes.

3 Justifications of Typological Diversity

We find that there is great variation in justifications for typological diversity claims. Some papers explain typological diversity through genealogy. For instance, Xu et al. (2022) *"select 24 typologically different languages covering a reasonable variety of language families"* and Zhang et al. (2023) create a dataset consisting of *"[18] languages that are both typologically close as well as distant from 10 language families and 13 sub-families"*.

Other papers use a selection of typological features, for instance, Mott et al. (2020) mention that *"the nine languages in our corpus cover five primary language families (...), and cover a range of morphological phenomena including suffixation, prefixation, (...)"*. Some papers also mention typological databases in their language selection, for instance, Gutierrez-Vasques et al. (2021) choose *"47 languages [from the] WALS 100-language sample, which aims to maximize both genealogical and areal diversity"*. Similarly, Muradoglu and Hulden (2022) consider *"typological diversity when selecting languages (...) [such as] languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS (...)"*. The most systematic approach to typologically diverse language selection we found is done by Jancso et al. (2020). They use a clustering algorithm on vectors with features from two typological databases to find the most distant clusters to sample languages from.

However, there is no consistent typological distance measurement for language selection. Thus, what is commonly meant by typological diverse language selection is not only inconsistent, but also often unsubstantiated.

4 Language Analysis

Next, we investigate the actual languages used in these datasets. Concretely, we aim to answer three questions: 1) how many languages are commonly used? 2) which languages are commonly used? and 3) how typologically diverse are these language selections?

First, we plot the number of languages the papers use in Figure 2. The number of languages

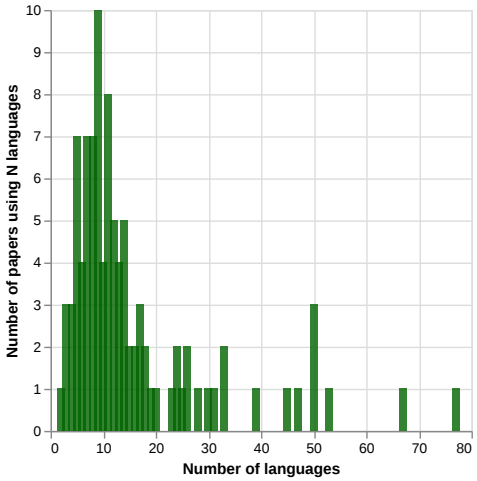

Figure 2: Number of papers using N languages.

used ranges from 2 to 77, with a mean of 16 and a standard deviation of 14. Four of the papers that contain a claim do not mention the languages they use. None the papers in our sample mention whether the number of languages they use relates to or is influenced by their typologically diversity claim. Similarly, only some papers, specifically ones that introduce a dataset, explicitly mention design choices with regard to the number of languages used.

Next, we look at the actual languages involved. The papers use 283 unique languages, of which 147 are used just once. English is the most-used language, followed by German, Finnish, Turkish, Russian and Spanish. Here, we observe a skew towards languages from the Eurasian macroarea.

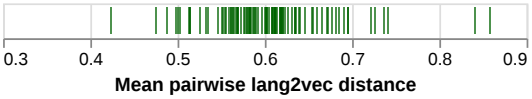

Figure 3: Mean pairwise syntactic lang2vec distance per paper.

Lastly, we approximate the actual typological diversity across papers by taking the average syntactic lang2vec (Littell et al., 2017) distance of all pairwise combinations in each paper's language set with coverage in lang2vec (97/103)[2]. The measured typological diversity varies across papers, with outliers on either side (Figure 3). The lowest mean pairwise distance (0.42) is found in Goel et al. (2022), who use *"3 typologically diverse languages – English, French and Spanish"*. The highest distance (0.86) is found in (Vania et al., 2019), who evaluate on North Sámi, Galician, and Kazah.

[2]Four papers do not mention the languages they use, two contain languages for which no ISO-693-3 code exists; Kholosi and Pomak.

References

Anmol Goel, Charu Sharma, and Ponnurangam Kumaraguru. 2022. An unsupervised, geometric and syntax-aware quantification of polysemy. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10565–10574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Anna Jancso, Steven Moran, and Sabine Stoll. 2020. The ACQDIV corpus database and aggregation pipeline. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 156–165, Marseille, France. European Language Resources Association.

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.

Justin Mott, Ann Bies, Stephanie Strassel, Jordan Kodner, Caitlin Richter, Hongzhi Xu, and Mitchell Marcus. 2020. Morphological segmentation for low resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3996–4002, Marseille, France. European Language Resources Association.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Jan Rijkhoff and Dik Bakker. 1998. Language sampling. *Linguistic Typology*, 2(3):263–314.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116, Hong Kong, China. Association for Computational Linguistics.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: How good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.