Wessel Poelman[♠*], Esther Ploeger[♦*], Miryam de Lhoneux[♠] & Johannes Bjerva[♦]
[♠]KU Leuven, Belgium [♦]Aalborg University, Denmark
wessel.poelman@kuleuven.be

A Call for Consistency in Reporting Typological Diversity

Claims about Multilingual NLP

- 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).
- Justifications for 'typological diversity' exhibit great variation; no set definition, methodology or consistent link to linguistic typology.

Findings

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.

Method

- Automatically search the entire ACL Anthology
-
```
typological.+?diverse|
typological.+?diversity|
diverse.+?typological
```
- Annotate if papers contain a claim (103/140)
- Two annotators, Cohen's κ = 0.64 (substantial)
- Approximate typological diversity using syntactic `lang2vec` distance (Littell et al., 2017)

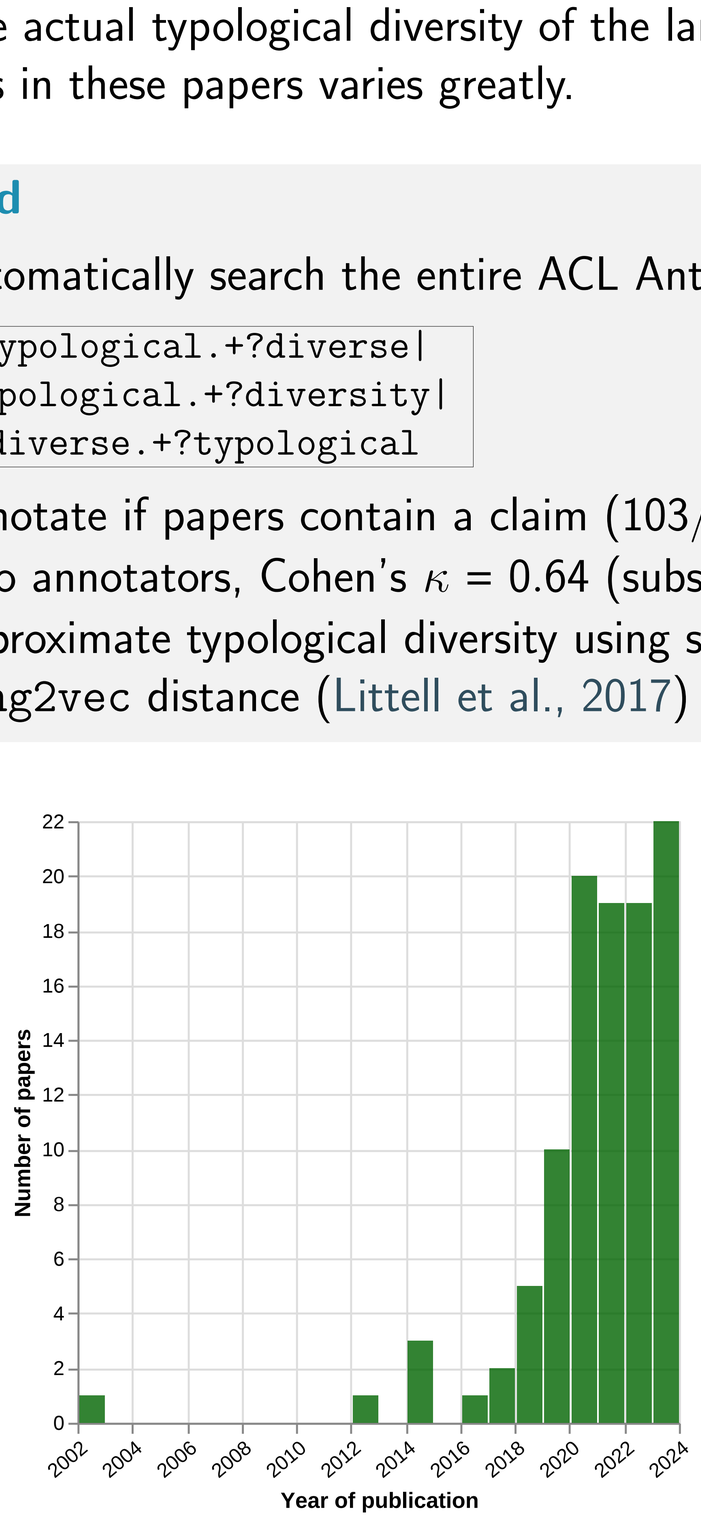

Figure 1: Number of papers in the ACL Anthology claiming a 'typologically diverse' set of languages over the years.

Recommendation

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

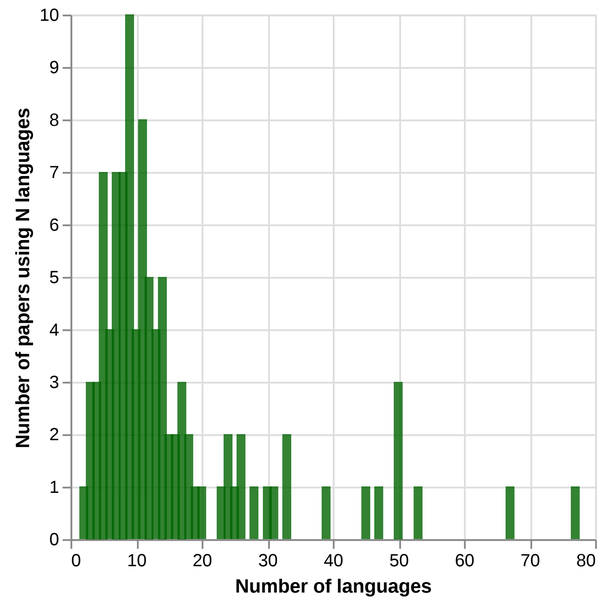

Figure 2: Number of papers using N languages. These range from 2 to 77 (mean 16, standard deviation 14). There are 283 unique languages, of which 147 are used just once (long tail).

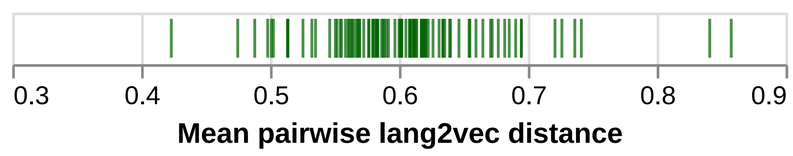

Figure 3: Mean pairwise syntactic lang2vec distance per paper (min 0.42, max 0.86).

Example Claims and Justifications

- Goel et al. (2022): *"3 typologically diverse languages – English, French and Spanish"*
- Vania et al. (2019): *"3 typologically diverse low-resource languages – North Sámi, Galician, and Kazah"*
- Xu et al. (2022): *"24 typologically different languages covering a reasonable variety of language families"*
- Zhang et al. (2023): *"[18] languages (...) both typologically close as well as distant from 10 language families and 13 sub-families"*.
- Mott et al. (2020): *"the 9 languages (...) cover five primary language families (...), and cover a range of morphological phenomena"*.
- Muradoglu and Hulden (2022): *"we consider typological diversity when selecting [30] languages (...) [such as] languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS"*.
- Jancso et al. (2020): use a clustering algorithm on vectors with features from two typological databases to find the most distant clusters to sample 14 languages from.

CARLSBERG FONDET EP & JB – This work was supported by a Semper Ardens: Accelerate research grant (CF21-0454) from the Carlsberg Foundation.
KU LEUVEN WP & ML – This work was funded by a KU Leuven BOF C1 grant (C14/23/096).