A Call for Consistency in Reporting Typological Diversity

SIGTYP (co-located with EACL 2024)

Wessel Poelman[*♣] Esther Ploeger[*✿] Miryam de Lhoneux[♣] Johannes Bjerva[✿]
wessel.poelman@kuleuven.be
espl@cs.aau.dk
[♣]KU Leuven, Belgium [✿]Aalborg University, Denmark

Multilingual NLP

Interest in multilingual NLP is increasing.

Multilingual NLP

Interest in multilingual NLP is increasing.

More and more work on generalizability *across languages*.

Multilingual NLP

Interest in multilingual NLP is increasing.

More and more work on generalizability *across languages*.

$\rightarrow$ Generalizability is increasingly claimed using *linguistic typology*.
"We evaluate on 12 typologically diverse languages."

Multilingual NLP

Interest in multilingual NLP is increasing.

More and more work on generalizability *across languages*.

$\rightarrow$ Generalizability is increasingly claimed using *linguistic typology*.
"We evaluate on 12 typologically diverse languages."

What does 'typologically diverse' even mean?

Data Collection

1. Collect papers from the ACL Anthology.
2. Annotate if they *claim* a 'typologically diverse' language set.
3. If yes, annotate which languages they use.

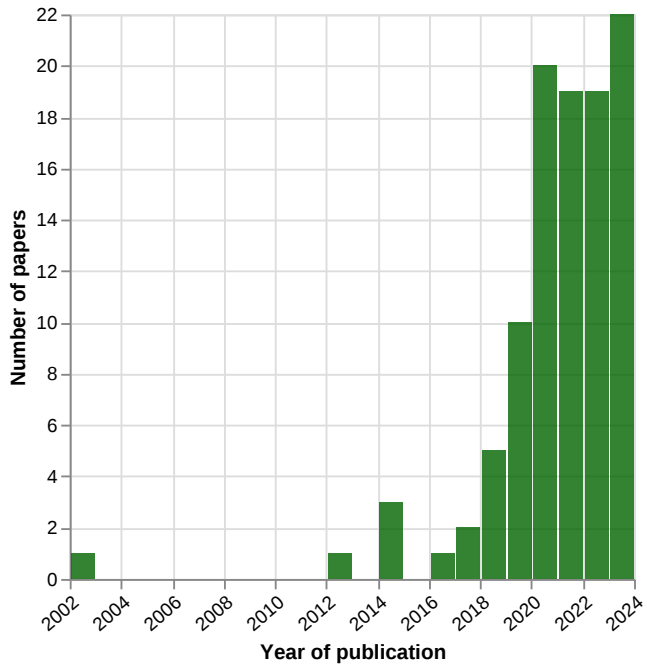
Data Collection

1. Collect papers from the ACL Anthology.
2. Annotate if they *claim* a 'typologically diverse' language set.
3. If yes, annotate which languages they use.

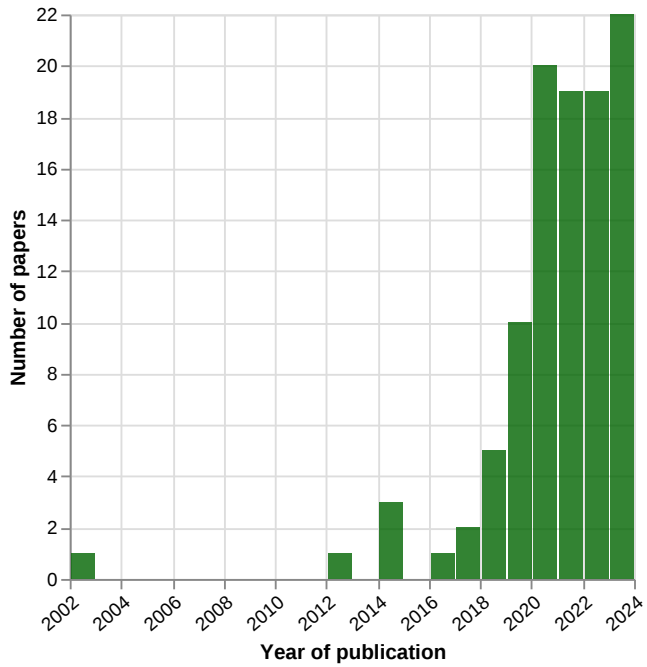
Annotation Results

- 140 papers total
- 103 paper contain a claim
- Cohens κ of 0.64

Usage

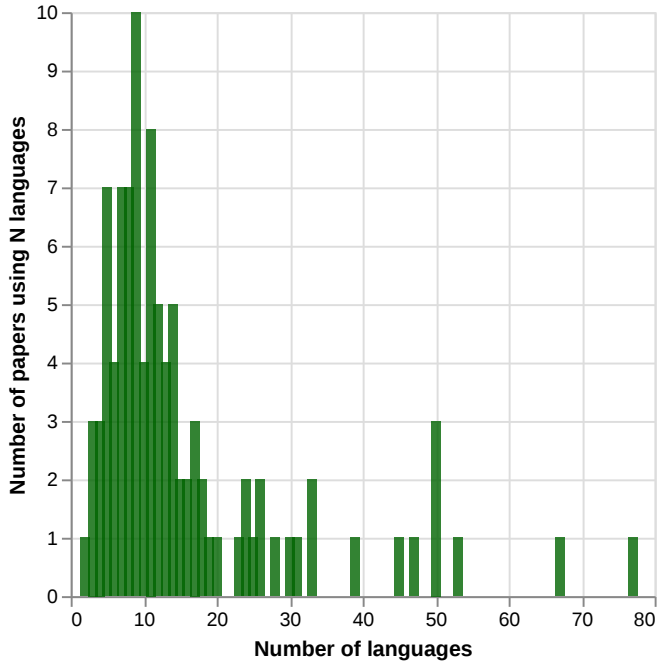

Number of papers
0
2
4
6
8
10
12
14
16
18
20
22
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Year of publication

Usage



- A quite recent trend.
- Claim occurrences are increasing.

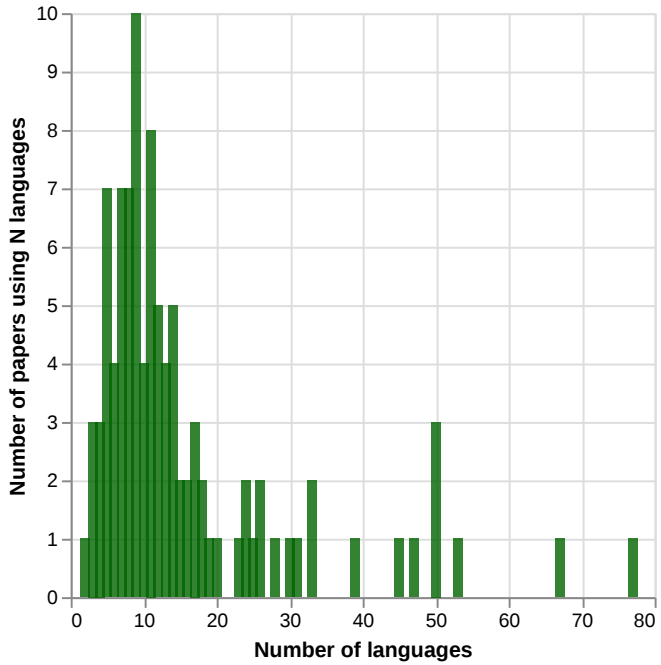
Number of Languages



AALBORG UNIVERSITY

KU LEUVEN

Number of Languages



- Number of languages used varies considerably (2 – 77).
- Most papers use between 5-20 languages.
- There are 283 unique languages, of which 147 are used once.

Justifications

- "24 typologically different languages covering a reasonable variety of language families"
- "[18] languages that are both typologically close as well as distant from 10 language families and 13 sub-families"
- "[30] languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS (. . .)"

Justifications

- "24 typologically different languages covering a reasonable variety of language families"
- "[18] languages that are both typologically close as well as distant from 10 language families and 13 sub-families"
- "[30] languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS (. . .)"

No consistency regarding number of languages, justifications or the relation between these, while using the same terminology.

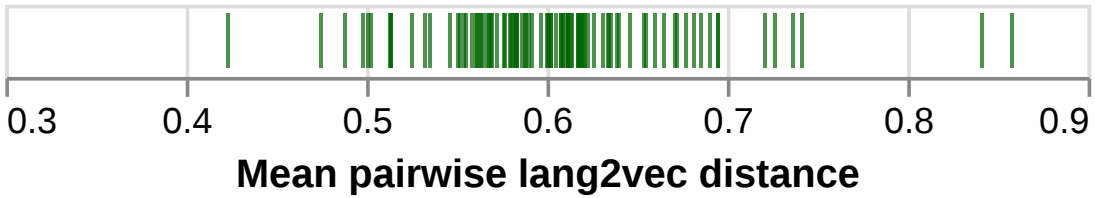
Justifications

- "24 typologically different languages covering a reasonable variety of language families"
- "[18] languages that are both typologically close as well as distant from 10 language families and 13 sub-families"
- "[30] languages that exhibit varying degrees of complexity for inflection. We also consider morphological characteristics coded in WALS (. . .)"

No consistency regarding number of languages, justifications or the relation between these, while using the same terminology.

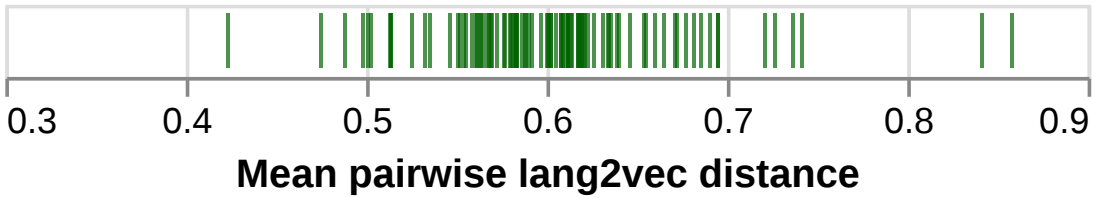
→ What if we approximate this?

Approximation



- Mean pairwise syntactic `lang2vec` distance per paper.
- Minimum of 0.42
  - English, French, and Spanish
- Maximum of 0.86
  - North Sámi, Galician, and Kazah

Approximation



- Mean pairwise syntactic `lang2vec` distance per paper.
- Minimum of 0.42
  - English, French, and Spanish
- Maximum of 0.86
  - North Sámi, Galician, and Kazah

Not ideal. . .

But it gives at least some approximation of what constitutes ‘typological diversity’.

Conclusion

Recommendation

1. Include an operationalization of 'typological diversity'.
   - Related to the phenomenon of interest.
   - Related to the number of languages used.
   - 'Why is our language selection typologically diverse?'
2. Ideally, show this using some empirical measure or approximation.

Final Remarks

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

Check out our (much more in-depth) pre-print about this:

What is 'Typological Diversity' in NLP?

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

Abstract

The NLP research community has devoted increased attention to languages beyond English, resulting in considerable improvements for multilingual NLP. However, these improvements only apply to a small subset of the world's languages. Aiming to extend this, an increasing number of papers aspires to enhance *generalizable* multilingual performance *across languages*. To this end, linguistic typology is commonly used to motivate language selection, on the basis that a broad typological sample ought to imply generalization across a broad range of languages. These selections are often described as being 'typologically diverse'. In this work, we systematically investigate NLP research that includes claims regarding 'typological diversity'. We find there are no set definitions or criteria for such claims. We introduce metrics to approximate the diversity of language selection along several axes and find that the results vary considerably across papers. Furthermore, we show that skewed language selection can lead to overestimated multilingual performance. We recommend future work to

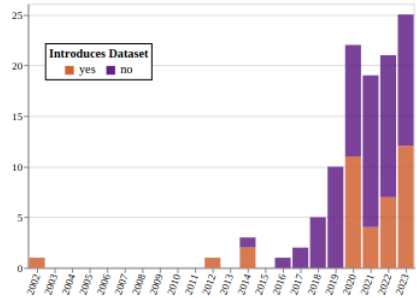


Figure 1: Number of papers with 'typological diversity' claims published by year.

Despite the potential of multilingual language modelling, common methodologies are primarily developed for English. But there is no guarantee that an approach that works well for one language will work equally well for others (Gerz et al., 2018). For instance, morphologically complex languages can be over-segmented by current widely-used tokenization methods (Rust et al., 2021). Evaluation

AALBORG UNIVERSITY KU LEUVEN