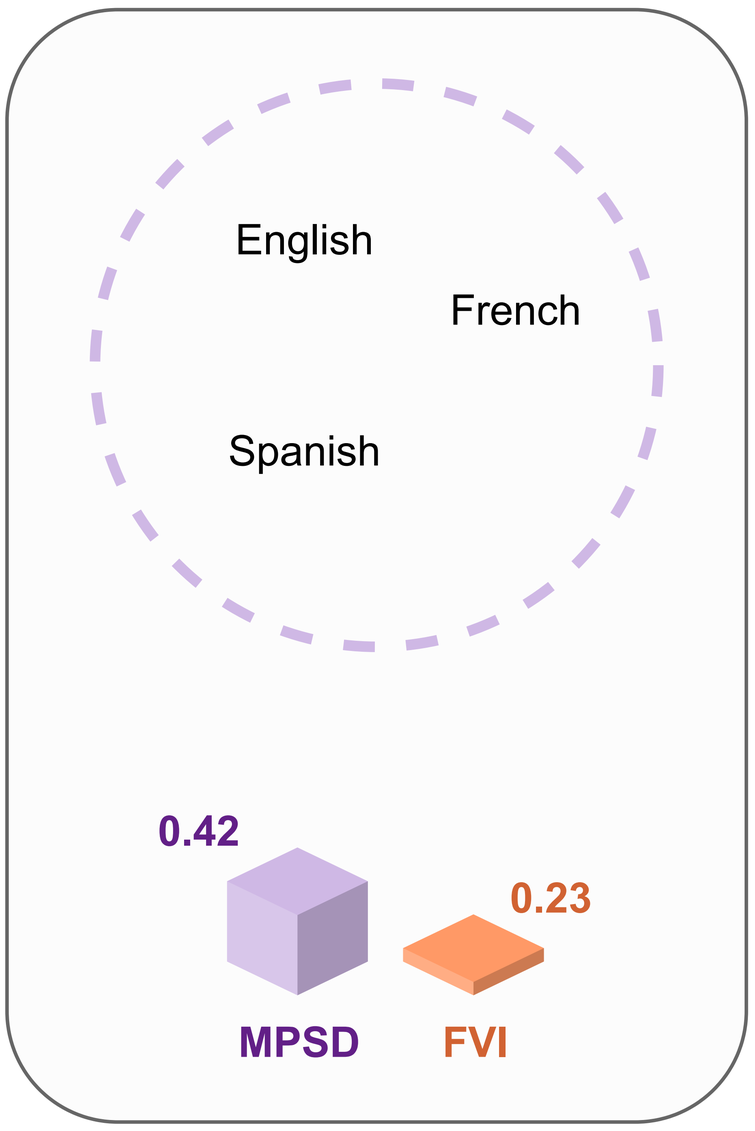
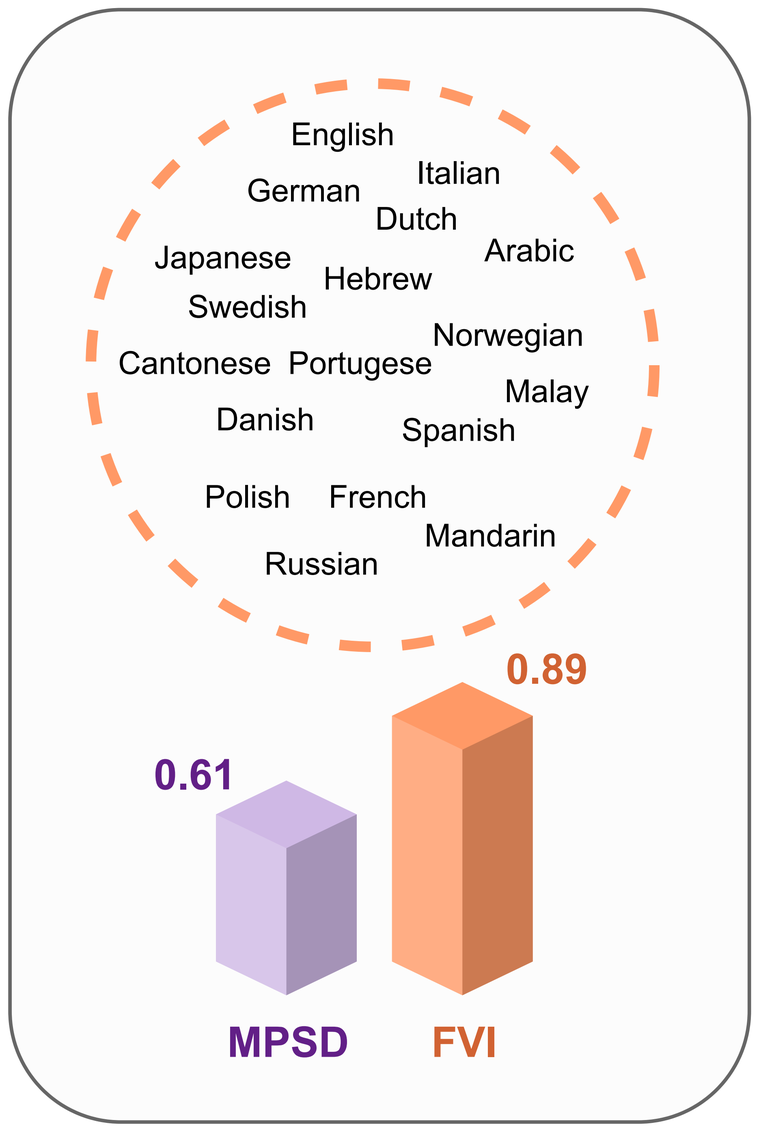
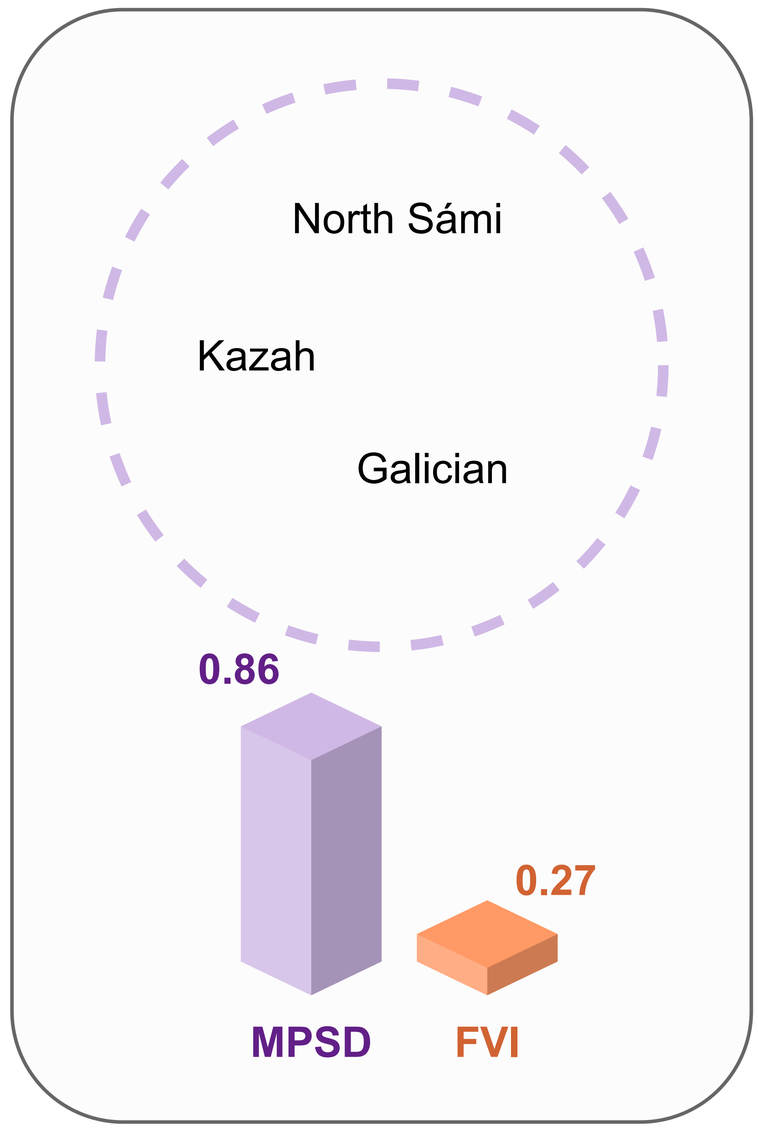
What is "Typological Diversity" in NLP?

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





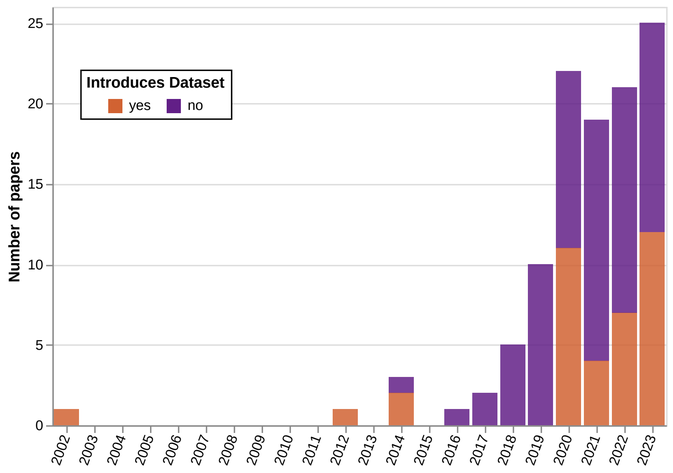


1. Background

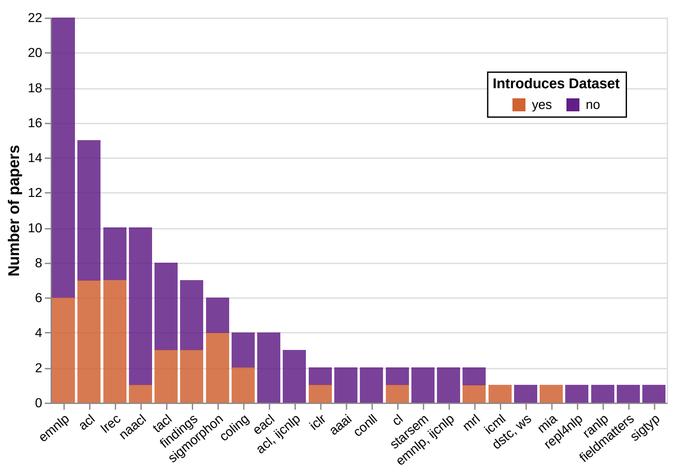
- **Multilinguality** is gaining interest in NLP.
- Some efforts try improving generalization *across* languages, loosely based on structural descriptions from **linguistic typology**. Increasingly, papers make claims of 'typologically diverse' language samples.
- However, it is not clear what is meant by these claims.

2. Contributions

1. A survey of the field for claims of 'typological diversity'.
2. Metrics to quantify diversity of language samples.
3. Insights into skewed language performance claims.



NLP and ML papers claiming to have 'typologically diverse' language samples.



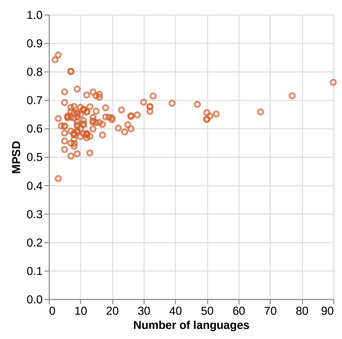
3. Justifications

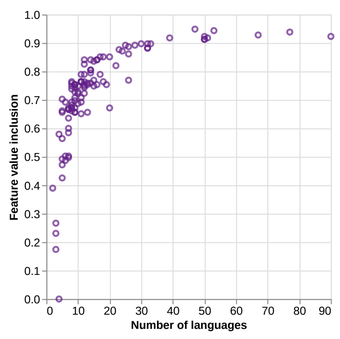
Paper	Justification	|L|
Xu et al. (2022)	"a reasonable variety of **language families**"	24
Muradoglu & Hulden (2022)	"varying degrees of complexity for **inflection** (...) also consider **morphological** characteristics coded in **WALS**"	30
Howell & Zamaraeva (2018)	"**genetically** and **geographically** diverse"	5

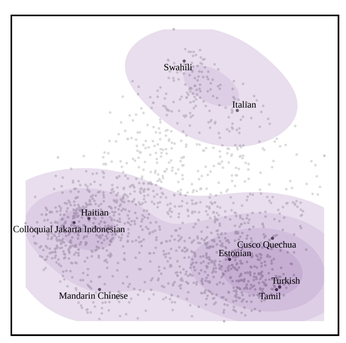
- Geography and Phylogeny are used as **proxies**.
- Number of languages used differs a lot.
- Justifications often not based on typology.

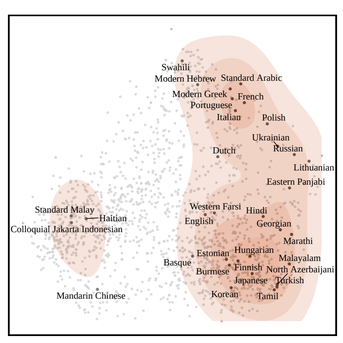
4. Metrics

Mean Pairwise Distance (MPD): `Geographic`, `Genetic`, `Syntactic` (MPSD) from URIEL → *Is the sample spread out?*
Feature Value Inclusion (FVI): Binary typological feature vectors from Grambank → *Are all values covered?*



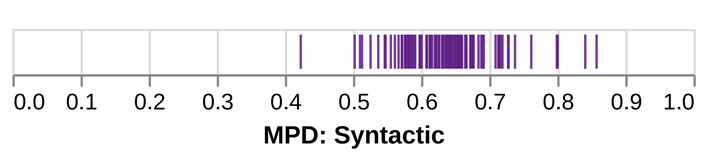


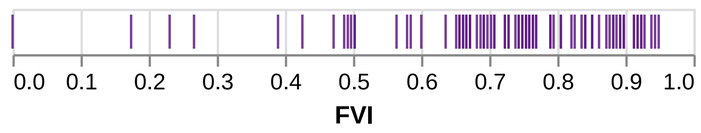


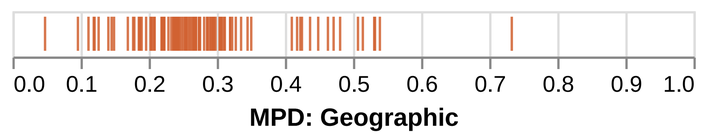


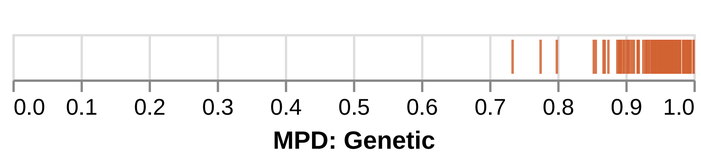
PCA of XCOPA and XTREME-R languages in Grambank.

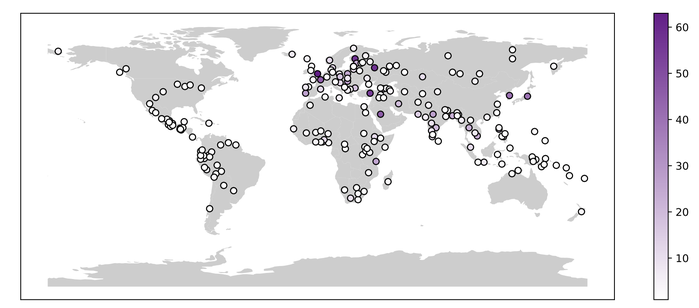
5. Approximations











6. Takeaways

1. Justify claims of 'typologically diverse' samples.
2. Phylogeny != Geography != Typology.
3. Check out the data and code, QR code below!

Acknowledgements

EP and JB are funded by the Carlsberg Foundation, under the *Semper Ardens: Accelerate* programme (project nr. CF21-0454). WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.

GitHub & Paper
LinkedIn