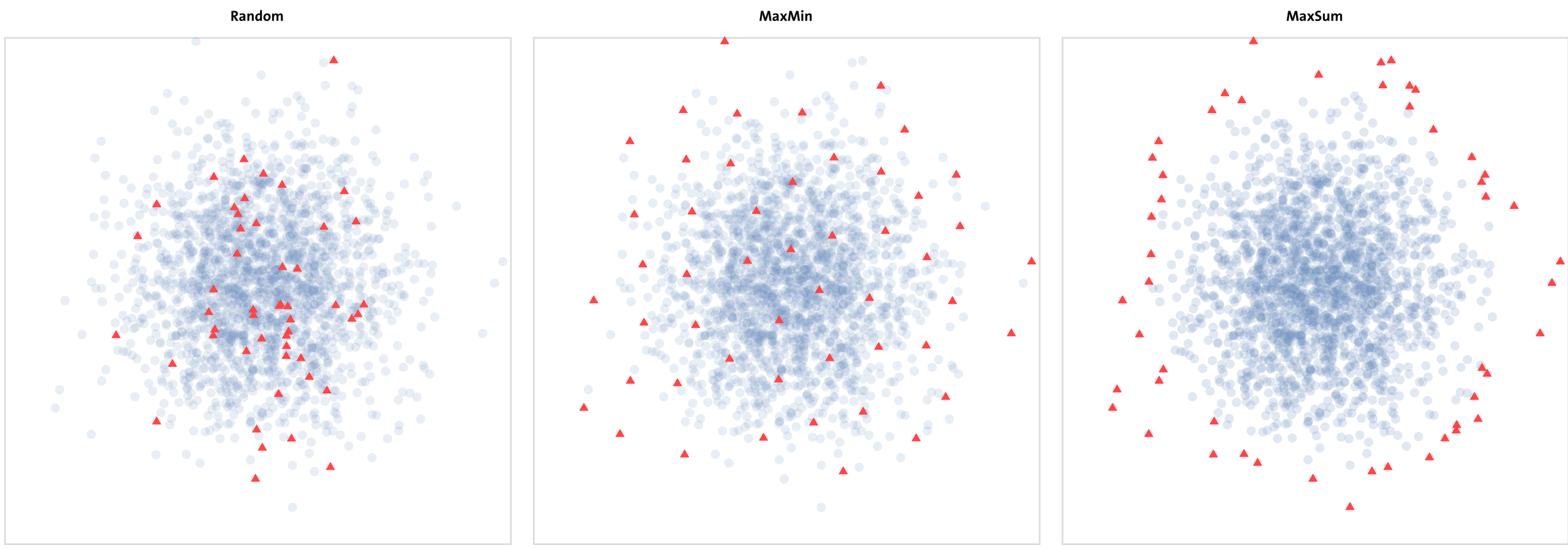
How Do I Choose Which Languages to Evaluate On?

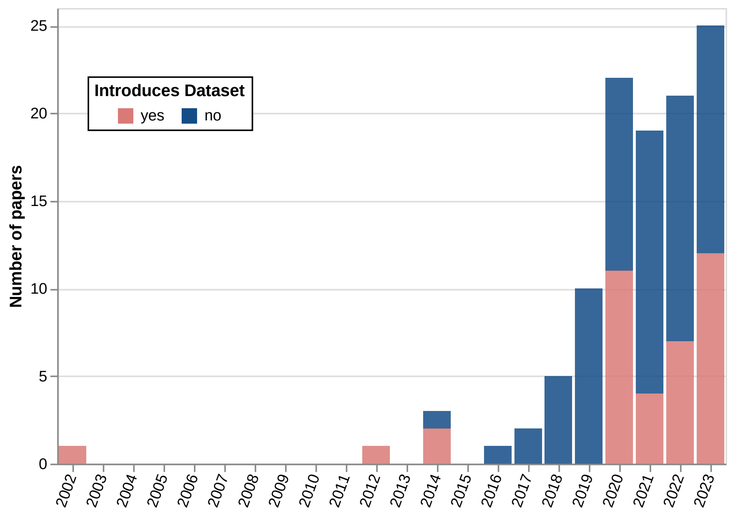
Esther Ploeger♢, Wessel Poelman♠, Andreas Holck Høeg-Petersen♢, Anders Schlichtkrull♢, Miryam de Lhoneux♠ & Johannes Bjerva♢

♢Aalborg University, Denmark ♠KU Leuven, Belgium wessel.poelman@kuleuven.be

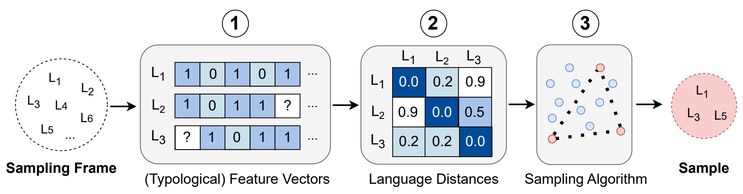


1. Background

- **Multilinguality** is gaining interest in NLP.
- Some efforts focus on improving generalization *across* languages, often loosely basing this on structural descriptions of languages from **linguistic typology**. An increasing number of papers make claims of 'typologically diverse' language samples.
- However, this link with linguistic typology is often vague and not principled, especially in **language sampling**.



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



Language sampling algorithms:

- **Random**: sample languages completely randomly
- **RandomFamily**: stratify by language family, sample uniformly and randomly
- **RandomGenus**: stratify by genus, sample uniformly and randomly
- **Convenience**: sample top k from most used languages in previous research
- **MaxSum**: sample most diverse → outliers (variety sampling in typology)
- **MaxMin**: sample most diverse → independence (probability sampling in typology)

2. Contributions

- A framework to systematically sample languages.
- Metrics to quantify linguistic diversity of language samples.
- Two sampling methods that select more diverse samples than random, convenience or phylogeny-inspired methods.

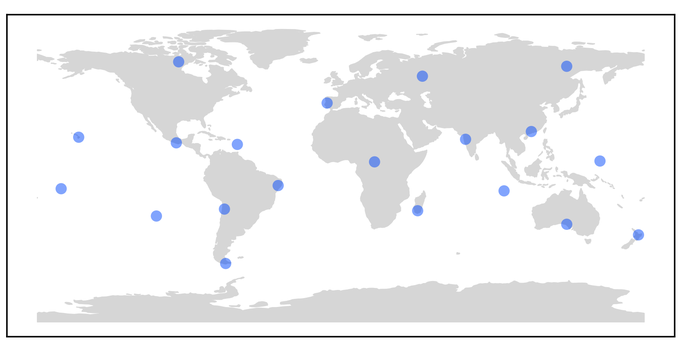
Metrics	
MPD:	Mean Pairwise Distance *Are we maximizing what we think?*
FVO:	Feature Value Overlap *Do we have overlap of values?*
FVI:	Feature Value Inclusion *Do we cover all feature values?*
$\mathcal{H}$:	Shannon Entropy *Is there spread in the feature values?*

3. Use Cases

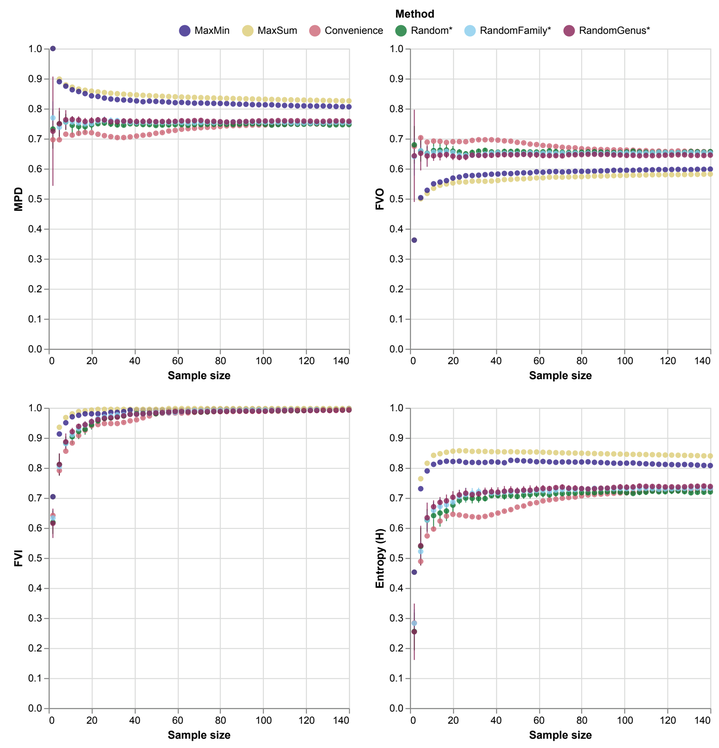
1. **Evaluation**: What is a good and diverse sample to test my phenomena of interest?
2. **Dataset expansion**: What are languages to add to my multilingual dataset to increase diversity or coverage?
3. **Other distance maximization**: Not just typological features, any language description works; What are the most *geographically* distant languages in my frame?

Effects in diversity metrics from adding Seri to UD v2.14.

MPD ↑	MPD' ↑	FVO ↓	FVO' ↓	FVI ↑	FVI' ↑	$\mathcal{H}$ ↑	$\mathcal{H}$' ↑
0.725	0.728	0.679	0.677	0.985	0.985	0.681	0.685



4. Intrinsic Evaluation



Methods with an asterisk are non-deterministic; scores are averages over 10 random runs and bars represent standard deviation.

5. Takeaways

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

Acknowledgements

This poster is based on:

- *What is "Typological Diversity" in NLP?*.
- *A Principled Framework for Evaluating on Typologically Diverse Languages*.

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.