What is "Typological Diversity" in NLP?

Esther Ploeger*
Department of Computer Science
Aalborg University

Wessel Poelman*
Department of Computer Science
KU Leuven

Miryam de Lhoneux
Department of Computer Science
KU Leuven

Johannes Bjerva
Department of Computer Science
Aalborg University

EMNLP – Nov 14, 2024

Are these language samples "typologically diverse"?

English
French
Spanish

English
Italian
German
Dutch
Japanese
Arabic
Hebrew
Swedish
Norwegian
Cantonese Portugese
Malay
Danish
Spanish
Polish French
Mandarin
Russian

North Sámi
Kazah
Galician

Are these language samples "typologically diverse"?

English
French

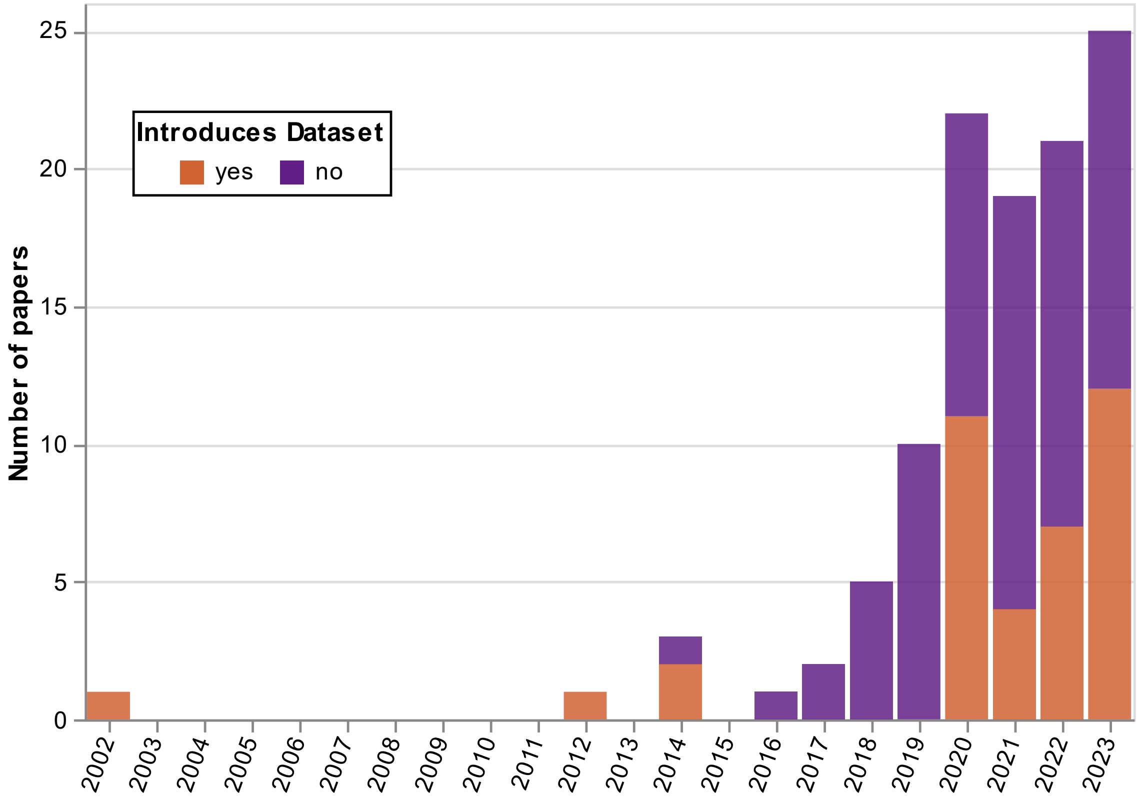
English
German
Italian
Dutch
Japanese
Arabic
Hebrew
Swedish
Norwegian

North Sámi

Maybe, but what does "typologically diverse" mean?

Multilingual NLP

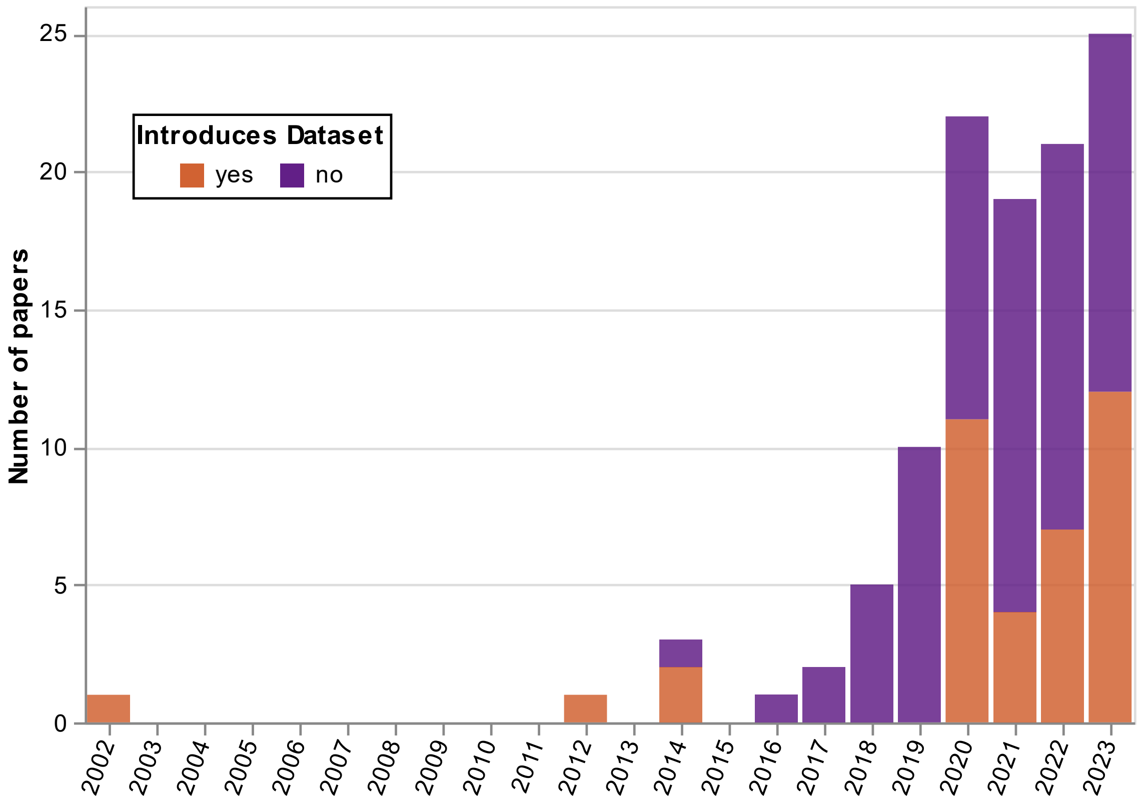
- Increased interest in **generalization** across languages.
- Loosely based on **linguistic typology**.
- *"We evaluate on a set of typologically diverse languages."*
- What does this mean?



Survey

```
typolog.+?div.+?|
div.+?typolog.+?
```

- ACL Anthology, NeurIPS, ICLR, ICML, AAAI & IJCAI
- Two annotators:
  - Claim?
  - Dataset
  - Languages



A brief look at language sampling in linguistic typology

- Create a sample that captures the **diversity** of the world's languages.
- Find **generalizations** across languages in the sample.
- Methods: *random*, *probablity, variety,* and *convenience*.

Rijkhoff, et al., 1993. A method of language sampling. Studies in Language.-- Velupillai. 2012. An Introduction to Linguistic Typology. -- Miestamo et al., 2016. Sampling for variety. Linguistic Typology. -- Rijkhoff & Bakker, 1998. Language sampling. Linguistic Typology.

A brief look at language sampling in linguistic typology

- Create a sample that captures the **diversity** of the world's languages.
- Find **generalizations** across languages in the sample.
- Methods: *random*, *probablity*, *variety*, and *convenience*.

Similar goals to multilingual NLP:
Generalization of models and datasets across languages

Rijkhoff, et al., 1993. A method of language sampling. Studies in Language.-- Velupillai. 2012. An Introduction to Linguistic Typology. -- Miestamo et al., 2016. Sampling for variety. Linguistic Typology. -- Rijkhoff & Bakker, 1998. Language sampling. Linguistic Typology.

A brief look at language sampling in linguistic typology

- Typology uses **geography** and **phylogeny** as priors for samples.
- NLP has access to the findings from typologists directly.
- Should we use the same priors...?

Rijkhoff, et al., 1993. A method of language sampling. Studies in Language.-- Velupillai. 2012. An Introduction to Linguistic Typology. -- Miestamo et al., 2016. Sampling for variety. Linguistic Typology. -- Rijkhoff & Bakker, 1998. Language sampling. Linguistic Typology.

A brief look at language sampling in linguistic typology

- Typology uses **geography** and **phylogeny** as priors for samples.
- NLP has access to the findings from typologists directly.
- Should we use the same priors…?

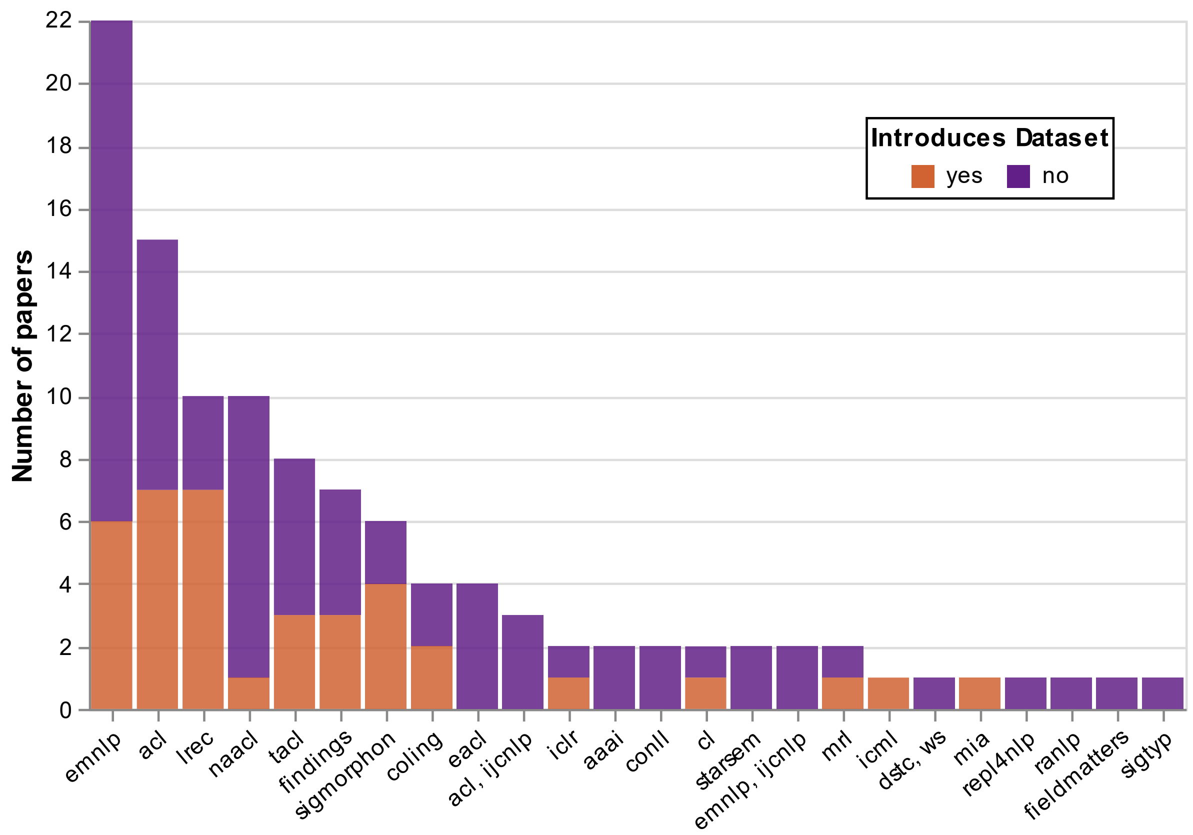
"The kind of variables that define **genealogical groups**
and **tree shapes** have a very **different** nature
from the kind of variables that define **typological diversity**."

- Stoll and Bickel (2013), based on Nichols (1996)

Stoll & Bickel, 2013. Capturing diversity in language acquisition research. Language Typology and Historical Contingency. -- Nichols, 1996. The comparative method as heuristic.

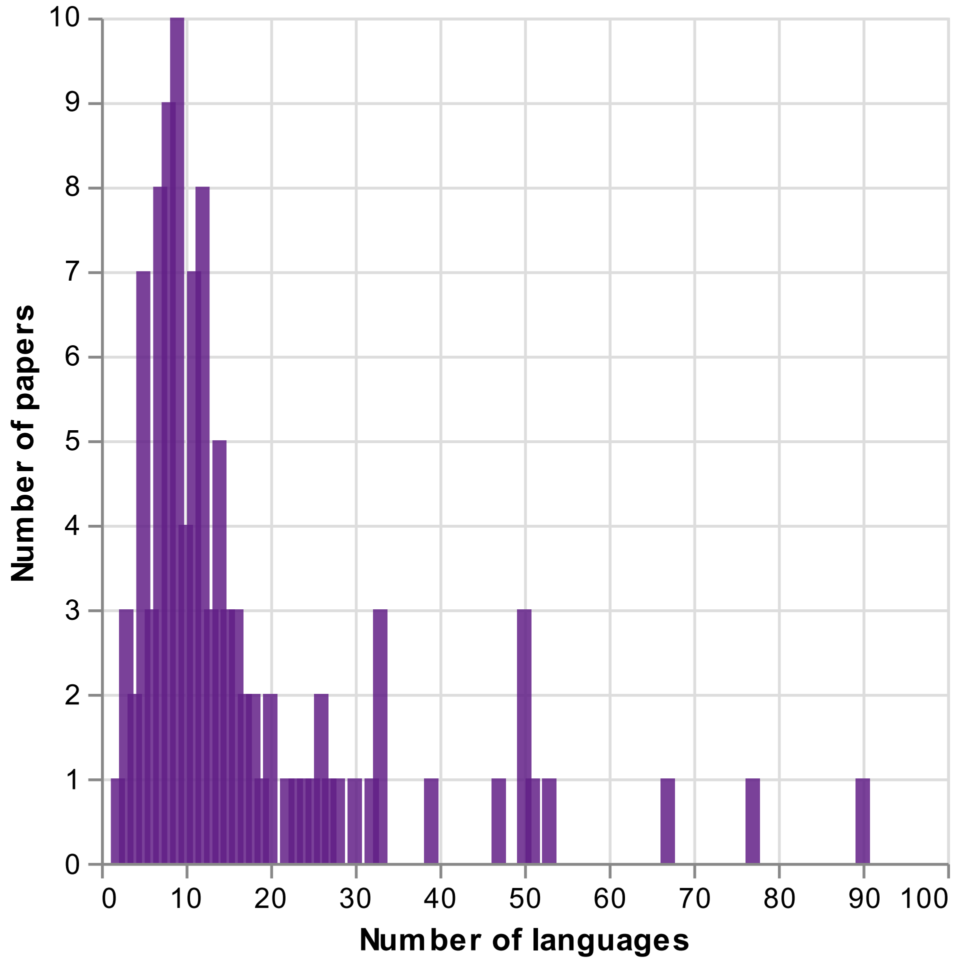
Collected Papers

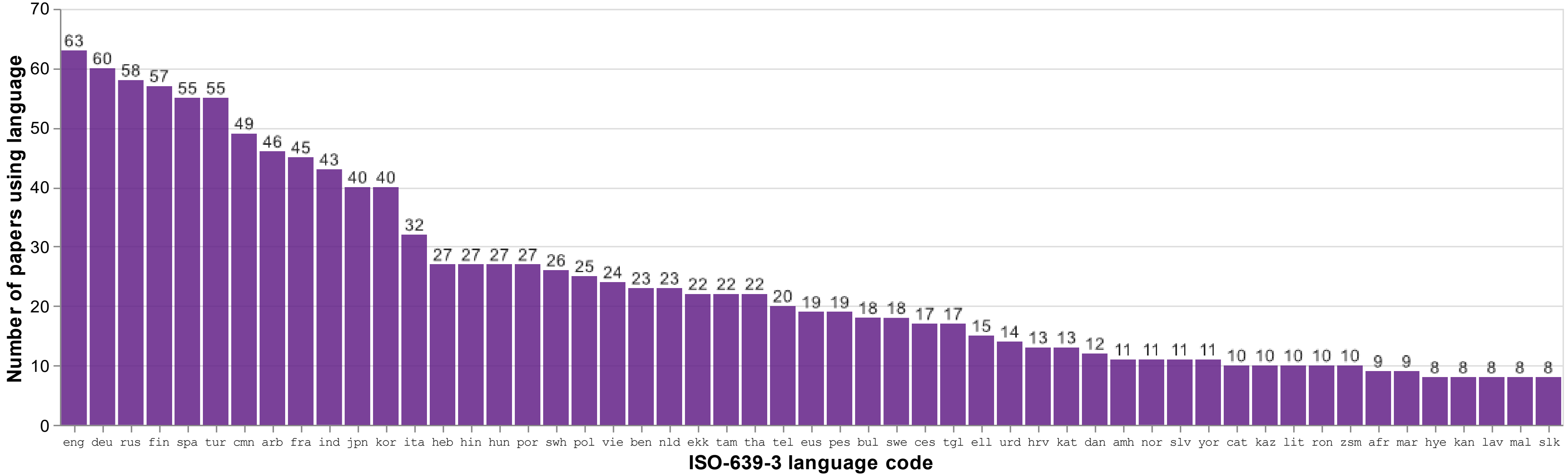
- 194 total, 110 with claim
- 38 introduce datasets
- Most at EMNLP, ACL, LREC, NAACL

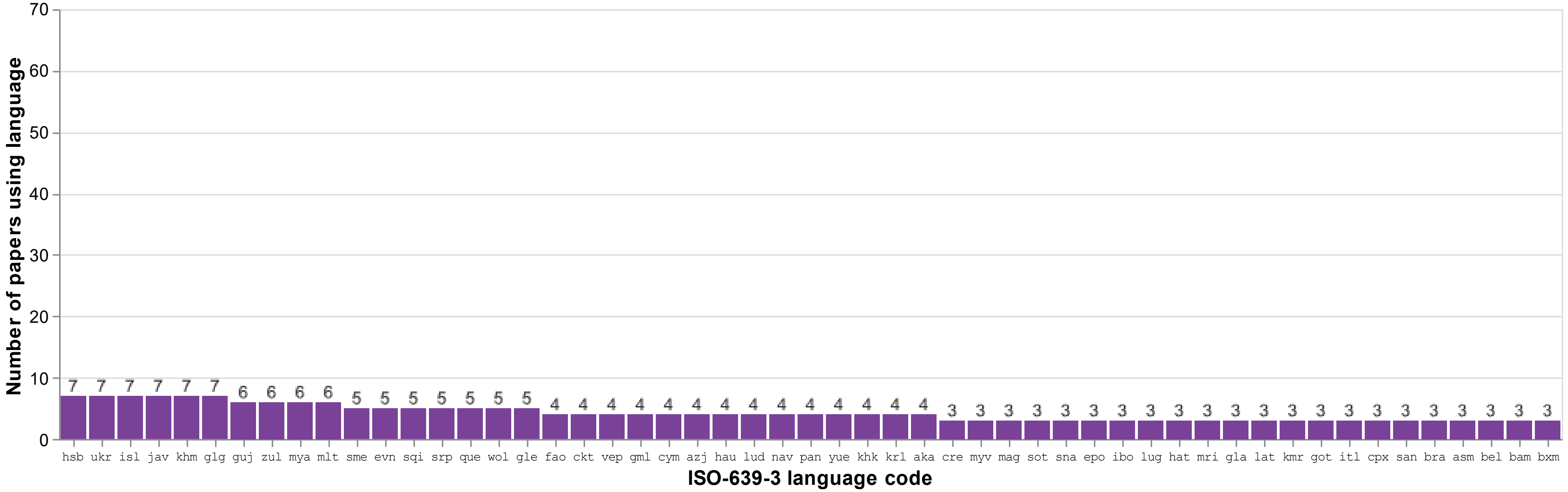


Languages

- 315 unique
- Range from 2 - 90 (median 11)
- 160 used just once
- 4 don't mention languages used
- Long tail of languages




Number of papers using language
ISO-639-3 language code
0 10 20 30 40 50 60 70
eng deu rus fin spa tur cmn arb fra ind jpn kor ita heb hin hun por swh pol vie ben nld ekk tam tha tel eus pes bul swe ces tgl ell urd hrv kat dan amh nor slv yor cat kaz lit ron zsm afr mar hye kan lav mal slk
63 60 58 57 55 55 49 46 45 43 40 40 32 27 27 27 27 26 25 24 23 23 22 22 22 20 19 19 18 18 17 17 15 14 13 13 12 11 11 11 11 10 10 10 10 10 9 9 8 8 8 8 8


Number of papers using language
ISO-639-3 language code
0 10 20 30 40 50 60 70
hsb ukr isl jav khm glg guj zul mya mlt sme evn sqi srp que wol gle fao ckt vep gml cym azj hau lud nav pan yue khk krl aka cre myv mag sot sna epo ibo lug hat mri gla lat kmr got itl cpx san bra asm bel bam bxm
7 7 7 7 7 7 6 6 6 6 5 5 5 5 5 5 5 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

Justifications from the papers	Number of languages used
"a reasonable variety of language **families**"	24
"languages from 10 language **families** and 13 **sub-families**"	18
"the languages in our corpus cover five primary language **families**, (. . .) and a range of **morphological phenomena**"	9
"languages that exhibit varying degrees of complexity for **inflection**. We also consider **morphological** characteristics coded in **WALS**"	30
"**genetically** and **geographically** diverse"	5

Can we approximate "typological diversity"?

- Many papers use **geography** and **phylogeny** as a **proxy**...
- However, **geography != phylogeny != typology**
- Use language descriptions to approximate.

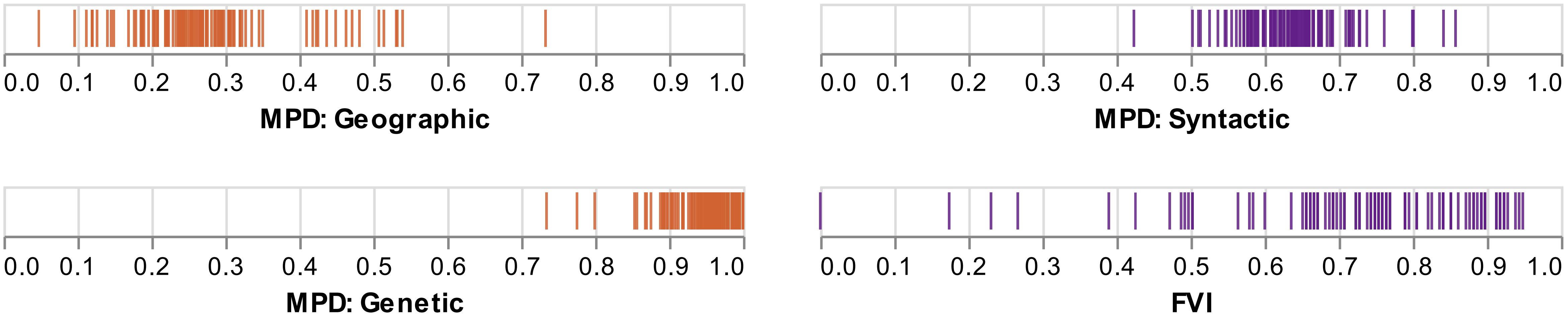
Can we approximate "typological diversity"?

- Many papers use **geography** and **phylogeny** as a **proxy**...
- However, **geography != phylogeny != typology**
- Use language descriptions to approximate.

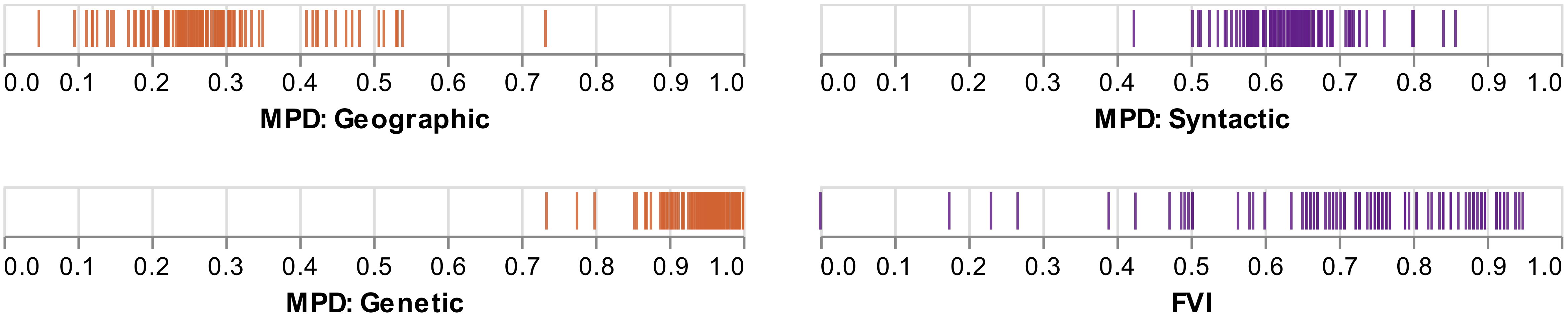
- Approximate the **proxies** using:
  - Geographic and genetic URIEL vectors
  - Calculate **Mean Pairwise Distance (MPD)** per sample
- Approximate **typological diversity** using:
  - Typological features from Grambank, calculate **Feature Value Inclusion (FVI)**
  - Syntactic URIEL vectors (using MPD, in graphs MPSD)

Miestamo et al., 2016. Sampling for variety. Linguistic Typology. -- Littell et al., 2017. URIEL and lang2vec. EACL. -- Wichmann & Holman, 2010. Pairwise comparisons of typological profiles. Rethinking Universals. -- Skirgård et al., 2023. Grambank

Approximations

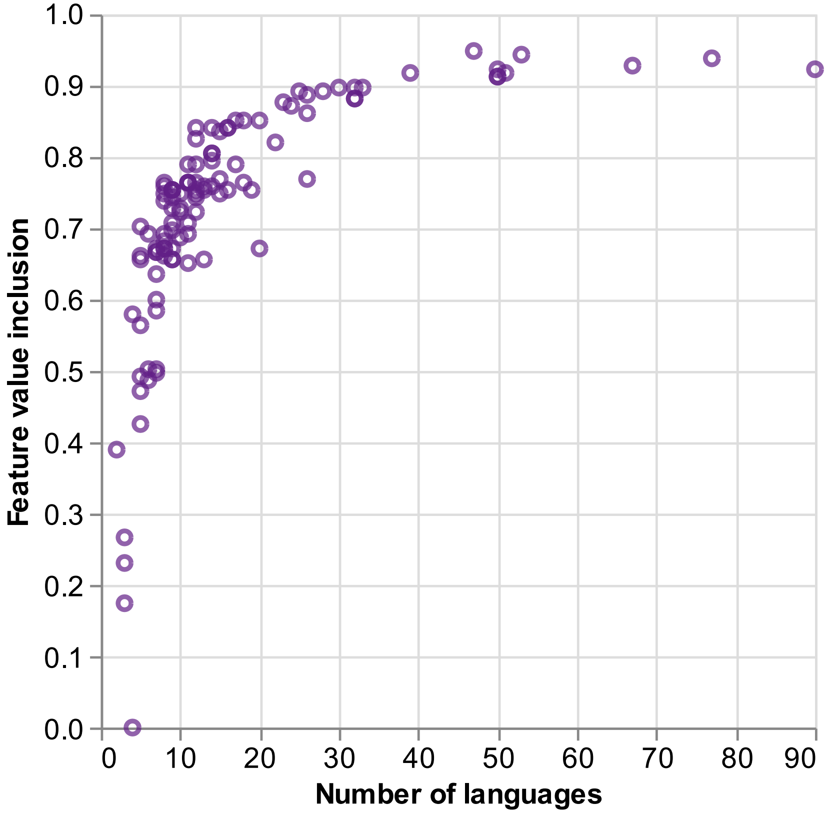


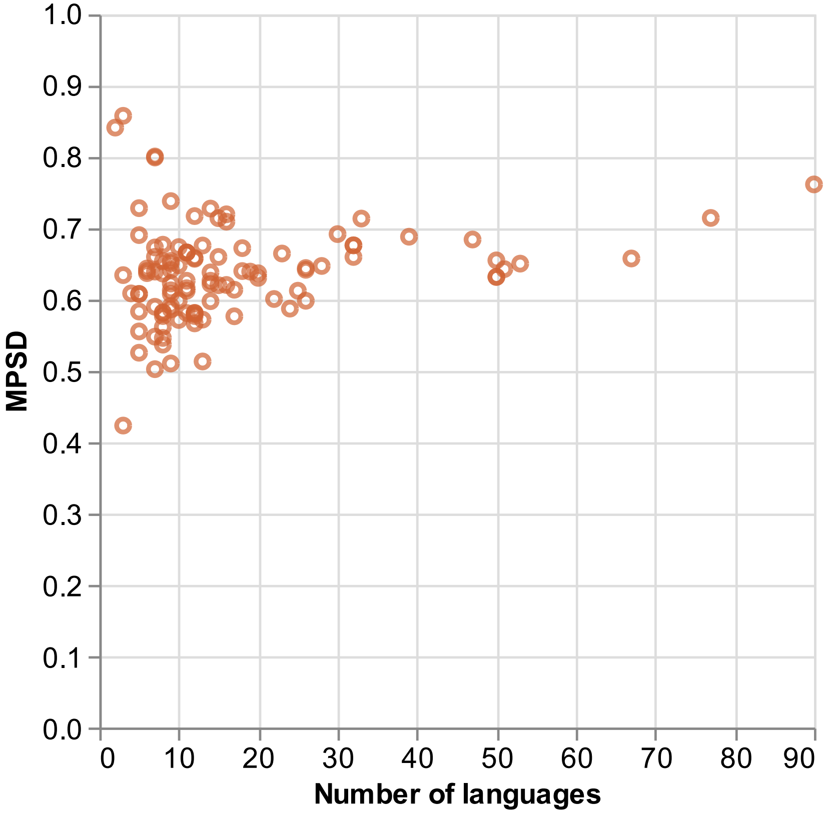
Approximations



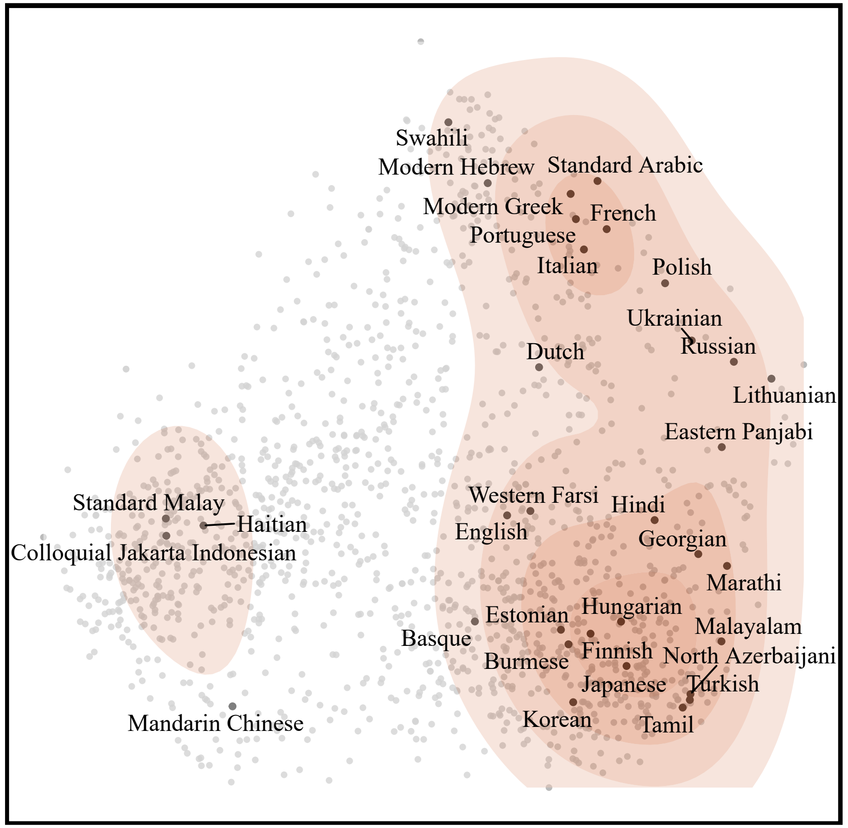
- A high **genetic** distance is achieved relatively quickly
- Feature Value Inclusion is very **spread out**
- What about the **number** of languages?

Approximations

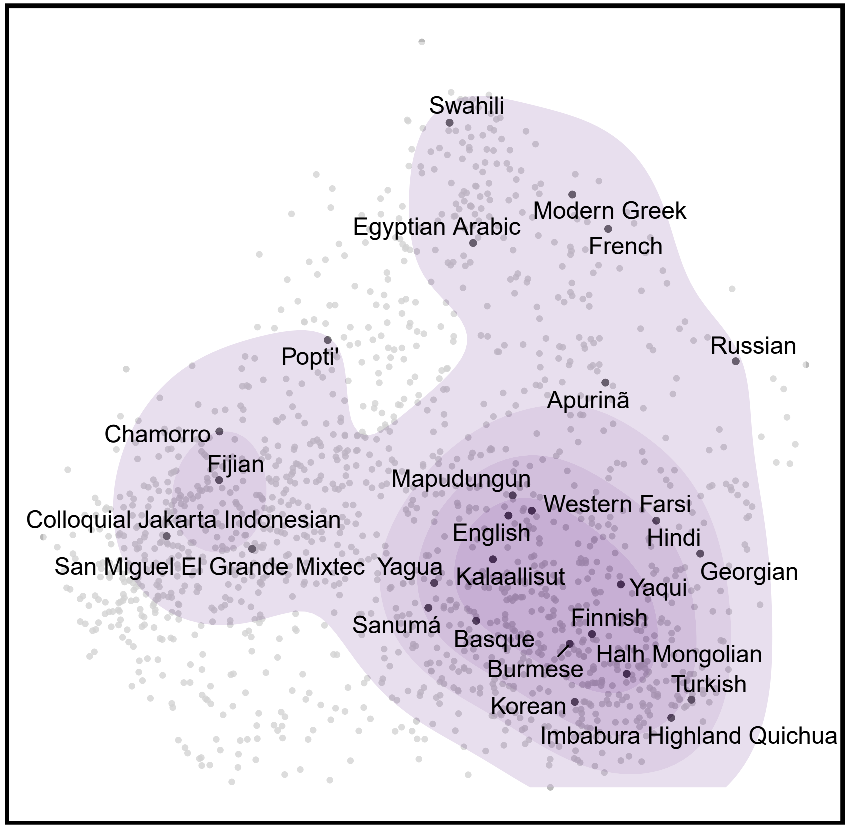




Another view



XTREME-R (0.92)



Paper with **highest FVI** (0.95)

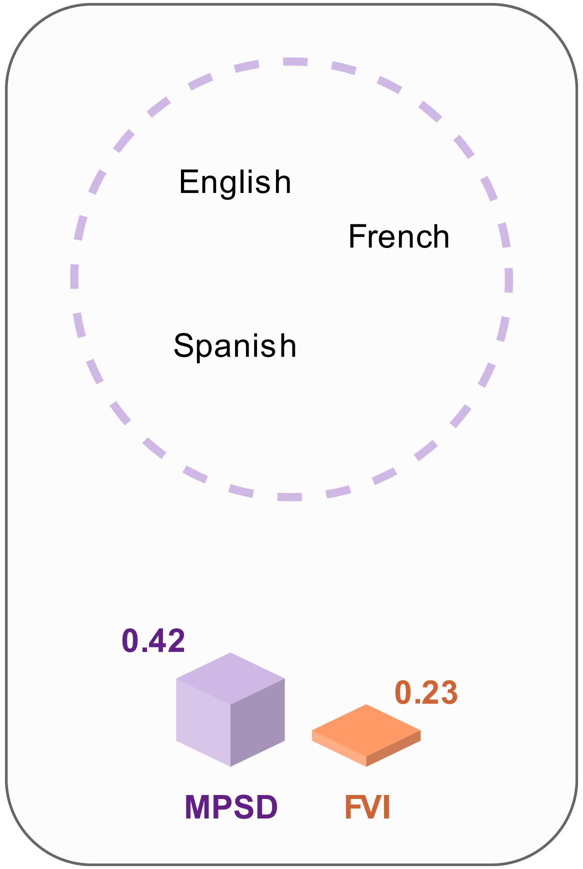
Are these language samples "typologically diverse"?

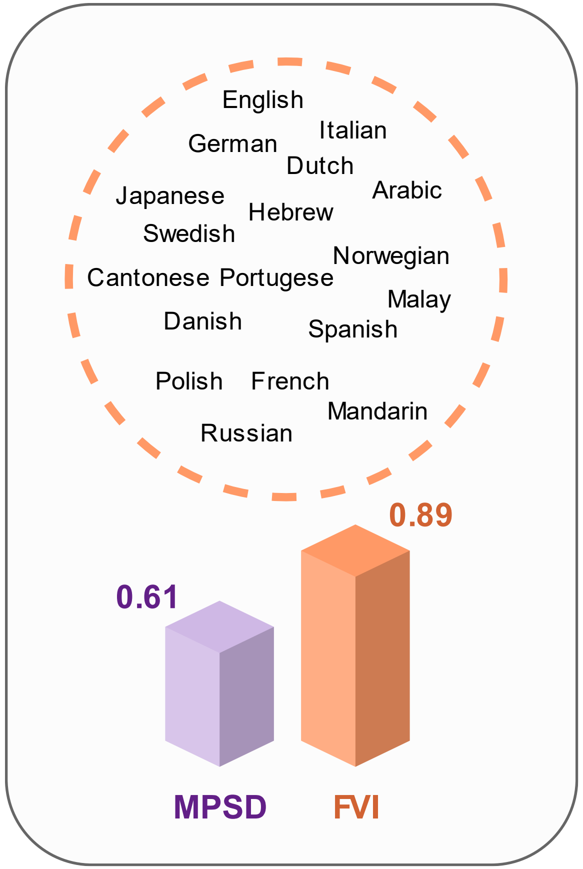
English
French
Spanish

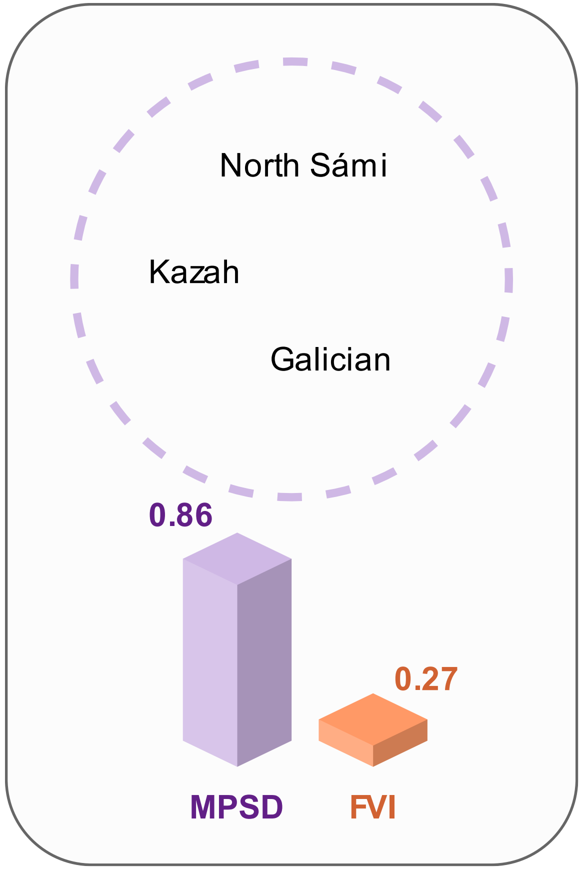
English
Italian
German
Dutch
Japanese
Arabic
Hebrew
Swedish
Norwegian
Cantonese Portugese
Malay
Danish
Spanish
Polish French
Mandarin
Russian

North Sámi
Kazah
Galician

Are these language samples "typologically diverse"?







Why does this matter?

- Datasets that **claim** to be typologically diverse spread.
- These claims set **expectations** regarding **generalization**.
- **Downstream** evaluations can be **skewed**:

Why does this matter?

- Datasets that **claim** to be typologically diverse spread.
- These claims set **expectations** regarding **generalization**.
- **Downstream** evaluations can be **skewed**:

"Be careful when reporting **averages** for multilingual benchmarks, especially if making **claims about multilinguality**."
- Anastasopoulos (2019)

"Using simple statistics, such as **average language performance**, might inject linguistic biases in favor of **dominant language families** into evaluation methodology."
- Pikuliak & Simko (2022)

Anastasopoulos. 2019. A note on evaluating multilingual benchmarks. Blog post. -- Pikuliak & Simko, 2022. Average Is Not Enough: Caveats of Multilingual Evaluation. MRL.

Case study: XTREME-R

"In order to catalyze meaningful progress, we extend XTREME to XTREME-R, which consists of an improved set of ten natural language understanding tasks, including **challenging language-agnostic** retrieval tasks, and covers **50 typologically diverse languages**."

- Ruder et al. (2021)

Case study: XTREME-R

"In order to catalyze meaningful progress, we extend XTREME to XTREME-R, which consists of an improved set of ten natural language understanding tasks, including **challenging language-agnostic** retrieval tasks, and covers **50 typologically diverse languages**."

- Ruder et al. (2021)

- **Not all languages** are available for all tasks.
- Performance is reported on as **average per task**.
- What about grouping by **typological** properties?

Case study: XTREME-R

Grouping languages per task by **inflection type**.

Highest number of languages in **orange**, lowest in **purple**.

Large gaps in **performance** and **coverage**.

Arguably not that diverse.

Subtask	Model	**Overall**	**By F**	Δ	Strong Pre	Weak Pre	Equal Pre & Suf	Strong Suf	Weak Suf	Little Aff	NA
Mewsli-X★	XLM-R-L	45.75 (*11*)	36.23 (*11*)	-9.52	- (*0*)	- (*0*)	- (*0*)	47.86 (*10*)	24.60 (*1*)	- (*0*)	- (*0*)
	mBERT	38.58 (*11*)	27.29 (*11*)	-11.29	- (*0*)	- (*0*)	- (*0*)	41.09 (*10*)	13.50 (*1*)	- (*0*)	- (*0*)
XNLI♦	XLM-R	79.24 (*15*)	76.54 (*15*)	-2.70	- (*0*)	71.20 (*1*)	- (*0*)	80.06 (*12*)	- (*0*)	78.35 (2)	- (*0*)
	mBERT	66.51 (*15*)	60.17 (*15*)	-6.35	- (*0*)	49.30 (*1*)	- (*0*)	68.60 (*12*)	- (*0*)	62.60 (2)	- (*0*)
	mT5	84.85 (*15*)	82.92 (*15*)	-1.92	- (*0*)	80.60 (*1*)	- (*0*)	85.57 (*12*)	- (*0*)	82.60 (2)	- (*0*)
LAReQA★	XLM-R-L	40.75 (*11*)	40.54 (*11*)	-0.22	- (*0*)	- (*0*)	- (*0*)	40.88 (9)	- (*0*)	40.20 (2)	- (*0*)
	mBERT	21.58 (*11*)	19.24 (*11*)	-2.35	- (*0*)	- (*0*)	- (*0*)	22.92 (9)	- (*0*)	15.55 (2)	- (*0*)
XQuAD♠	XLM-R-L	77.21 (*11*)	77.24 (*11*)	+0.04	- (*0*)	- (*0*)	- (*0*)	77.19 (9)	- (*0*)	77.30 (2)	- (*0*)
	mBERT	65.05 (*11*)	61.84 (*11*)	-3.21	- (*0*)	- (*0*)	- (*0*)	66.89 (9)	- (*0*)	56.80 (2)	- (*0*)
	mT5	81.54 (*11*)	80.55 (*11*)	-0.99	- (*0*)	- (*0*)	- (*0*)	82.10 (9)	- (*0*)	79.00 (2)	- (*0*)
MLQA♠	XLM-R-L	72.71 (7)	73.33 (7)	+0.62	- (*0*)	- (*0*)	- (*0*)	72.47 (6)	- (*0*)	74.20 (*1*)	- (*0*)
	mBERT	61.30 (7)	60.84 (7)	-0.46	- (*0*)	- (*0*)	- (*0*)	61.48 (6)	- (*0*)	60.20 (*1*)	- (*0*)
	mT5	75.59 (7)	75.97 (7)	+0.38	- (*0*)	- (*0*)	- (*0*)	75.43 (6)	- (*0*)	76.50 (*1*)	- (*0*)
Tatoeba♦	XLM-R	77.29 (*41*)	64.92 (*36*)	-12.36	- (*0*)	31.30 (*1*)	58.60 (*1*)	82.10 (*28*)	76.37 (*3*)	77.43 (*3*)	63.74 (*5*)
	mBERT	43.33 (*41*)	32.03 (*36*)	-11.30	- (*0*)	12.10 (*1*)	31.00 (*1*)	49.24 (*28*)	39.27 (*3*)	32.90 (*3*)	27.68 (*5*)
UD-POS♠	XLM-R-L	74.96 (38)	71.12 (36)	-3.84	- (*0*)	- (*0*)	74.30 (*1*)	79.75 (28)	71.05 (2)	45.98 (5)	84.50 (2)
	mBERT	70.90 (38)	64.43 (36)	-6.47	- (*0*)	- (*0*)	59.30 (*1*)	75.51 (28)	60.75 (2)	48.66 (5)	77.95 (2)
XCOPA♦	XLM-R	69.22 (*11*)	65.93 (9)	-3.28	- (*0*)	61.80 (*1*)	- (*0*)	73.93 (6)	- (*0*)	75.30 (2)	52.70 (2)
	mBERT	56.05 (*11*)	54.75 (9)	-1.30	- (*0*)	52.20 (*1*)	- (*0*)	57.70 (6)	- (*0*)	56.20 (2)	52.90 (2)
	mT5	74.89 (*11*)	73.24 (9)	-1.65	- (*0*)	74.10 (*1*)	- (*0*)	78.00 (6)	- (*0*)	77.60 (2)	63.25 (2)
WikiANN-NER♠	XLM-R-L	64.43 (*48*)	62.02 (*40*)	-2.41	- (*0*)	69.90 (*1*)	62.10 (*1*)	66.92 (*31*)	61.37 (*3*)	48.17 (*4*)	63.66 (*8*)
	mBERT	62.68 (*48*)	61.73 (*40*)	-0.95	- (*0*)	72.70 (*1*)	65.00 (*1*)	64.93 (*31*)	57.23 (*3*)	49.38 (*4*)	61.12 (*8*)
TyDiQA♠	XLM-R-L	64.29 (9)	62.57 (8)	-1.72	- (*0*)	66.40 (*1*)	- (*0*)	65.67 (6)	- (*0*)	59.10 (*1*)	59.10 (*1*)
	mBERT	58.36 (9)	55.09 (8)	-3.26	- (*0*)	59.70 (*1*)	- (*0*)	60.97 (6)	- (*0*)	46.20 (*1*)	53.50 (*1*)
	mT5	81.94 (9)	83.73 (8)	+1.78	- (*0*)	87.20 (*1*)	- (*0*)	80.52 (6)	- (*0*)	83.60 (*1*)	83.60 (*1*)

Ruder et al., 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. EMNLP. -- Dryer, 2013. Prefixing vs. suffixing in inflectional morphology. WALS (v2020.3).

Case study: XTREME-R

Grouping languages per task by **word order**.

Highest number of languages in orange, lowest in **purple**.

Large gaps in **performance** and **coverage**.

Arguably not that diverse.

Subtask	Model	**Overall**	**By F**	Δ	OSV	OVS	VOS	SVO	SOV	VSO	NDO	NA
MLQA♠	XLM-R-L	72.71 (7)	70.83 (7)	-1.88	- (0)	- (0)	- (0)	75.22 (4)	70.80 (1)	67.00 (1)	70.30 (1)	- (0)
	mBERT	61.30 (7)	57.14 (7)	-4.16	- (0)	- (0)	- (0)	66.85 (4)	49.90 (1)	51.60 (1)	60.20 (1)	- (0)
	mT5	75.59 (7)	74.06 (7)	-1.53	- (0)	- (0)	- (0)	77.62 (4)	75.30 (1)	70.20 (1)	73.10 (1)	- (0)
LAReQA★	XLM-R-L	40.75 (11)	39.31 (11)	-1.44	- (0)	- (0)	- (0)	42.10 (6)	39.75 (2)	34.60 (1)	40.80 (2)	- (0)
	mBERT	21.58 (11)	19.75 (11)	-1.83	- (0)	- (0)	- (0)	24.10 (6)	15.10 (2)	17.00 (1)	22.80 (2)	- (0)
TyDiQA♠	XLM-R-L	64.29 (9)	62.80 (9)	-1.49	- (0)	- (0)	- (0)	67.26 (5)	58.55 (2)	62.60 (2)	- (0)	- (0)
	mBERT	58.36 (9)	56.91 (9)	-1.44	- (0)	- (0)	- (0)	61.24 (5)	55.55 (2)	53.95 (2)	- (0)	- (0)
	mT5	81.94 (9)	81.45 (9)	-0.50	- (0)	- (0)	- (0)	82.94 (5)	78.40 (2)	83.00 (2)	- (0)	- (0)
XQuAD♠	XLM-R-L	77.21 (11)	77.11 (11)	-0.10	- (0)	- (0)	- (0)	76.70 (6)	76.55 (2)	74.40 (1)	80.80 (2)	- (0)
	mBERT	65.05 (11)	63.55 (11)	-1.50	- (0)	- (0)	- (0)	67.35 (6)	56.60 (2)	62.20 (1)	68.05 (2)	- (0)
	mT5	81.54 (11)	81.04 (11)	-0.49	- (0)	- (0)	- (0)	82.22 (6)	79.10 (2)	80.30 (1)	82.55 (2)	- (0)
XNLI♦	XLM-R	79.24 (15)	78.57 (15)	-0.67	- (0)	- (0)	- (0)	80.31 (9)	75.10 (3)	77.20 (1)	81.65 (2)	- (0)
	mBERT	66.51 (15)	65.79 (15)	-0.72	- (0)	- (0)	- (0)	68.19 (9)	60.03 (3)	66.00 (1)	68.95 (2)	- (0)
	mT5	84.85 (15)	84.71 (15)	-0.14	- (0)	- (0)	- (0)	85.39 (9)	81.83 (3)	84.50 (1)	87.10 (2)	- (0)
Mewsli-X★	XLM-R-L	45.75 (11)	45.66 (11)	-0.09	- (0)	- (0)	- (0)	53.16 (5)	35.98 (4)	28.70 (1)	64.80 (1)	- (0)
	mBERT	38.58 (11)	37.88 (11)	-0.71	- (0)	- (0)	- (0)	47.28 (5)	27.93 (4)	15.30 (1)	61.00 (1)	- (0)
Tatoeba♦	XLM-R	77.29 (41)	72.82 (38)	-4.46	- (0)	- (0)	- (0)	81.42 (18)	75.74 (14)	64.55 (2)	86.57 (4)	55.83 (3)
	mBERT	43.33 (41)	39.66 (38)	-3.67	- (0)	- (0)	- (0)	52.57 (18)	33.04 (14)	25.10 (2)	54.78 (4)	32.83 (3)
XCOPA♦	XLM-R	69.22 (11)	66.16 (9)	-3.06	- (0)	- (0)	- (0)	72.89 (7)	72.90 (2)	- (0)	- (0)	52.70 (2)
	mBERT	56.05 (11)	55.17 (9)	-0.88	- (0)	- (0)	- (0)	57.11 (7)	55.50 (2)	- (0)	- (0)	52.90 (2)
	mT5	74.89 (11)	72.62 (9)	-2.27	- (0)	- (0)	- (0)	77.61 (7)	77.00 (2)	- (0)	- (0)	63.25 (2)
WikiANN-NER♠	XLM-R-L	64.43 (48)	65.10 (43)	+0.67	- (0)	- (0)	- (0)	66.80 (20)	59.74 (17)	57.95 (2)	79.70 (4)	61.30 (5)
	mBERT	62.68 (48)	63.98 (43)	+1.30	- (0)	- (0)	- (0)	67.34 (20)	54.26 (17)	58.80 (2)	75.92 (4)	63.58 (5)
UD-POS♠	XLM-R-L	74.96 (38)	77.45 (36)	+2.50	- (0)	- (0)	- (0)	74.59 (20)	69.65 (10)	71.55 (2)	86.97 (4)	84.50 (2)
	mBERT	70.90 (38)	71.60 (36)	+0.69	- (0)	- (0)	- (0)	72.74 (20)	62.82 (10)	61.25 (2)	83.22 (4)	77.95 (2)

Limitations

- Coverage: no typological database covers every aspect of every language.
- Survey is based on abstracts and titles, papers may contain claims in other sections.
- Phylogeny and geography are useful for NLP, but arguably not for making claims about typological diversity.

Limitations

- Coverage: no typological database covers every aspect of every language.
- Survey is based on abstracts and titles, papers may contain claims in other sections.
- Phylogeny and geography are useful for NLP, but arguably not for making claims about typological diversity.

We believe that the reporting of typological diversity can be more principled than it currently is, despite incomplete resources!

Summary

1. There is **no consistent** definition or methodology when making ‘typological diversity’ claims.

2. Our approximations of typological diversity exhibit **considerable variation** across papers.

3. Averages and aggregated results can give **distorted** views of multilingual performance estimates.

Recommendations

1. Include an **operationalization** of 'typological diversity'.
2. Possibly add some **empirical justification**.
3. Including these has the potential to benefit multilingual NLP by allowing more **fine-grained insights**.

Acknowledgements

Paper

Thanks to:

- The TypNLP group at Aalborg University for useful feedback (especially Heather Lent).
- The LAGoM-NLP group at KU Leuven for insightful discussions and feedback (Thomas Bauwens for helping to create the tables using fiject).
- The audience for listening!

LinkedIn

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

wessel.poelman@kuleuven.be & espl@cs.aau.dk