Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter

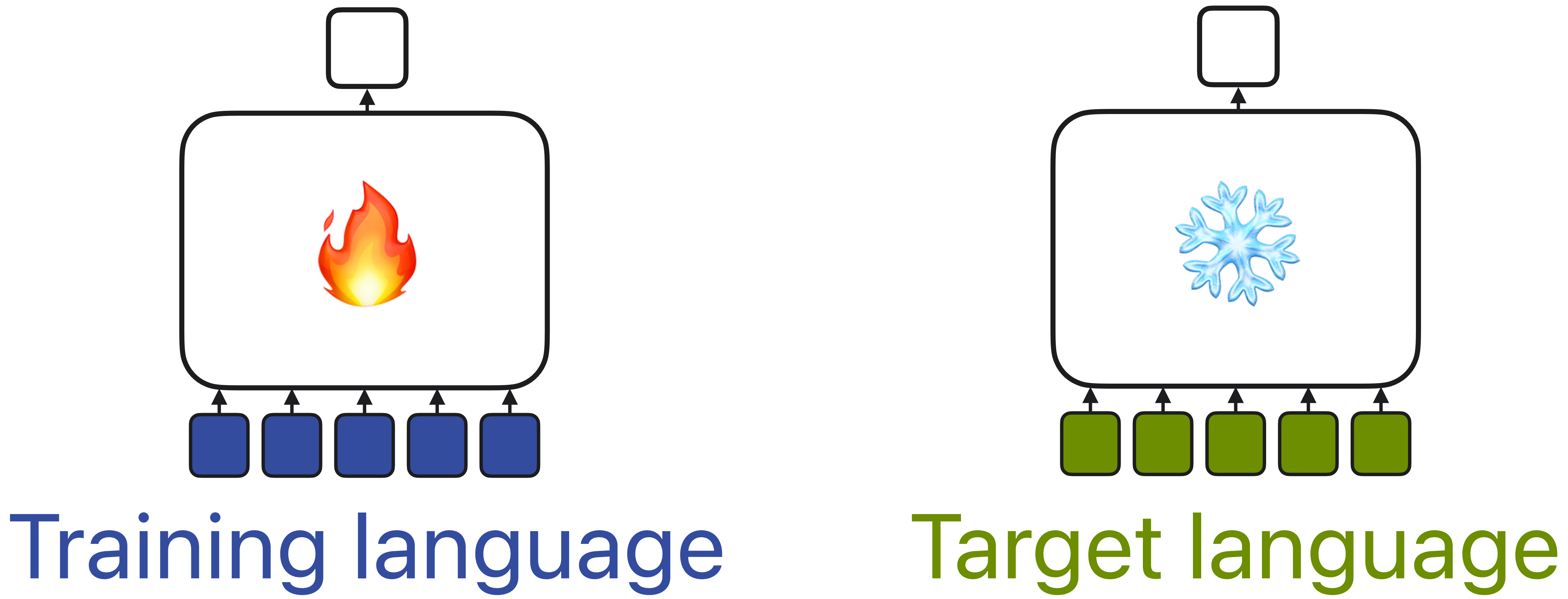
Verena Blaschke,[1] Masha Fedzechkina,[2] Maartje ter Hoeve[2]

ACL Findings | [1]LMU Munich & MCML, [2]Apple | July 2025

Motivation

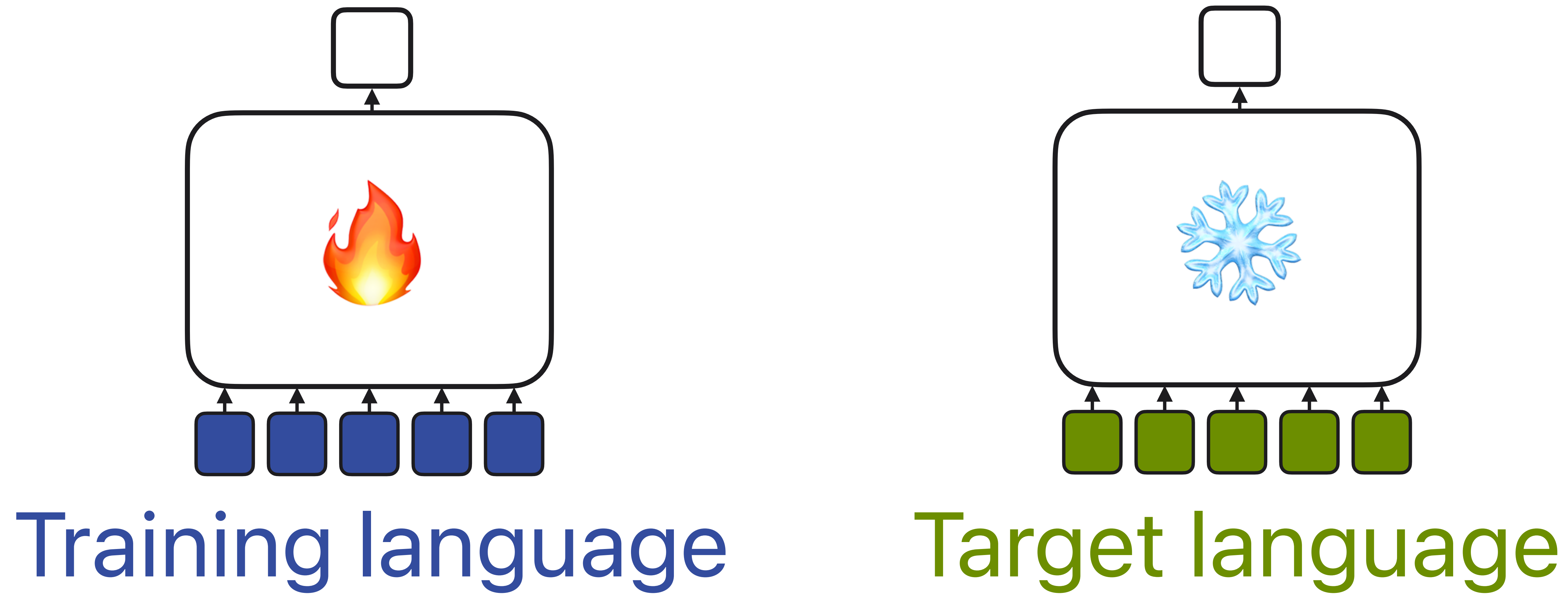
- 7000+ languages in the world
  - Only a few of which are focused on in NLP research (Joshi+, 2020)
    - Only a few of which have training data available

→ Cross-lingual transfer



Motivation

- Given a target language, how do I select a good training language?
  - Intuitively: pick a language/dataset that is *in some way* similar…
    … but in *what way*?
  - Prior work: either relatively few languages or NLP tasks (Philippy+, 2023)
  - Here: 263 languages, 3 NLP tasks, 10 similarity measures



Overview

Large-scale cross-lingual transfer experiments

Correlations with similarity measures

Practical takeaways for picking source languages

Overview

Large-scale cross-lingual transfer experiments

Correlations with similarity measures

Practical takeaways for picking source languages

NLP experiments

POS tagging & dep. parsing

Topic classification

NLP experiments — grammatical tasks

POS tagging & dep. parsing

- Universal Dependencies
- 70 × 153 languages
- UDPipe 2 (mono- and multilingual char/word embeddings)

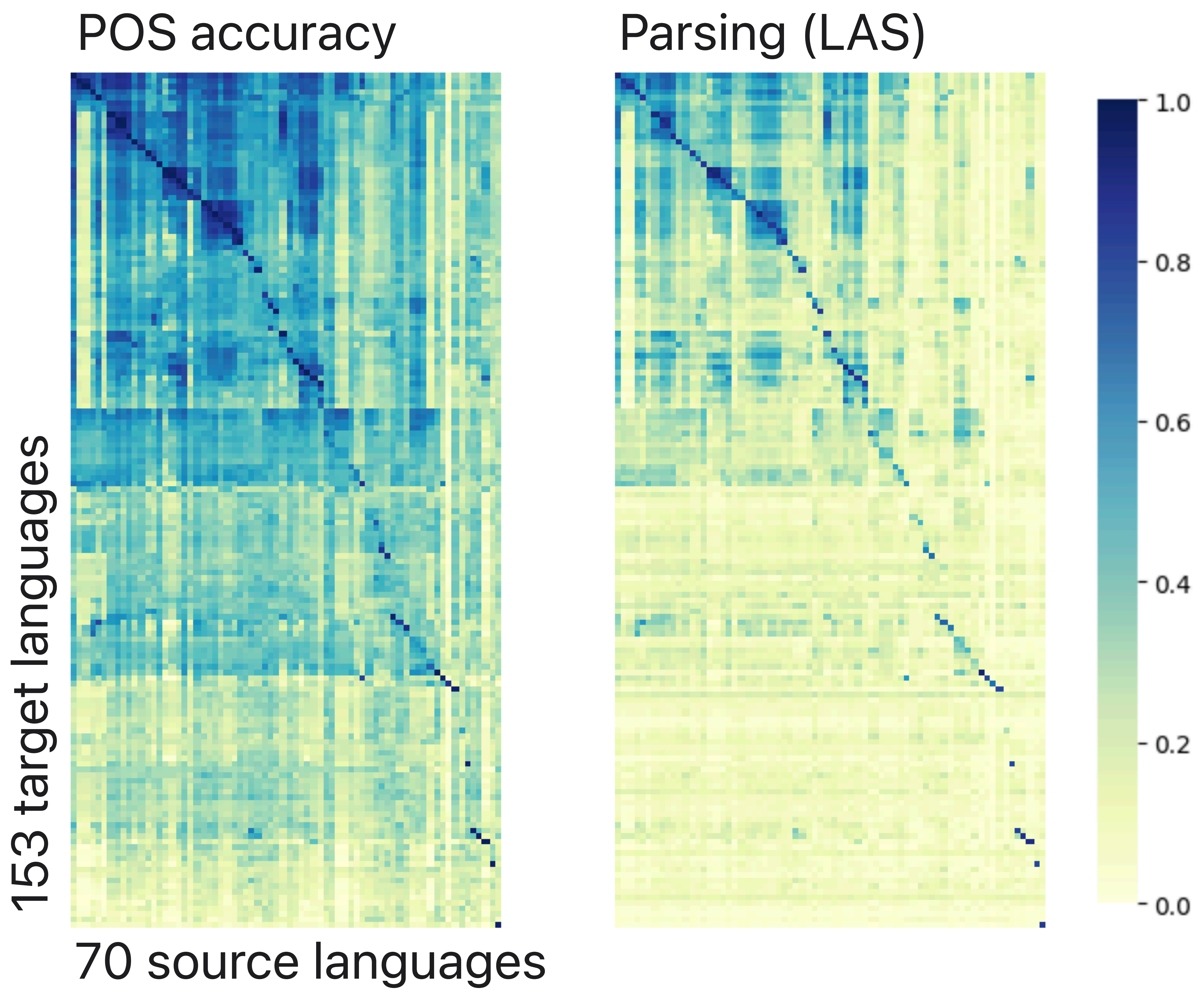
Topic classification

Universal Dependencies (de Marneffe et al., CL 2021)
UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task (Straka, CoNLL 2018)

NLP experiments — grammatical tasks

POS tagging & dep. parsing

- Universal Dependencies
- 70 × 153 languages
- UDPipe 2 (mono- and multilingual char/word embeddings)



LAS = labelled attachment score

Universal Dependencies (de Marneffe et al., CL 2021)
UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task (Straka, CoNLL 2018)

NLP Experiments — topic classification

POS tagging & dep. parsing

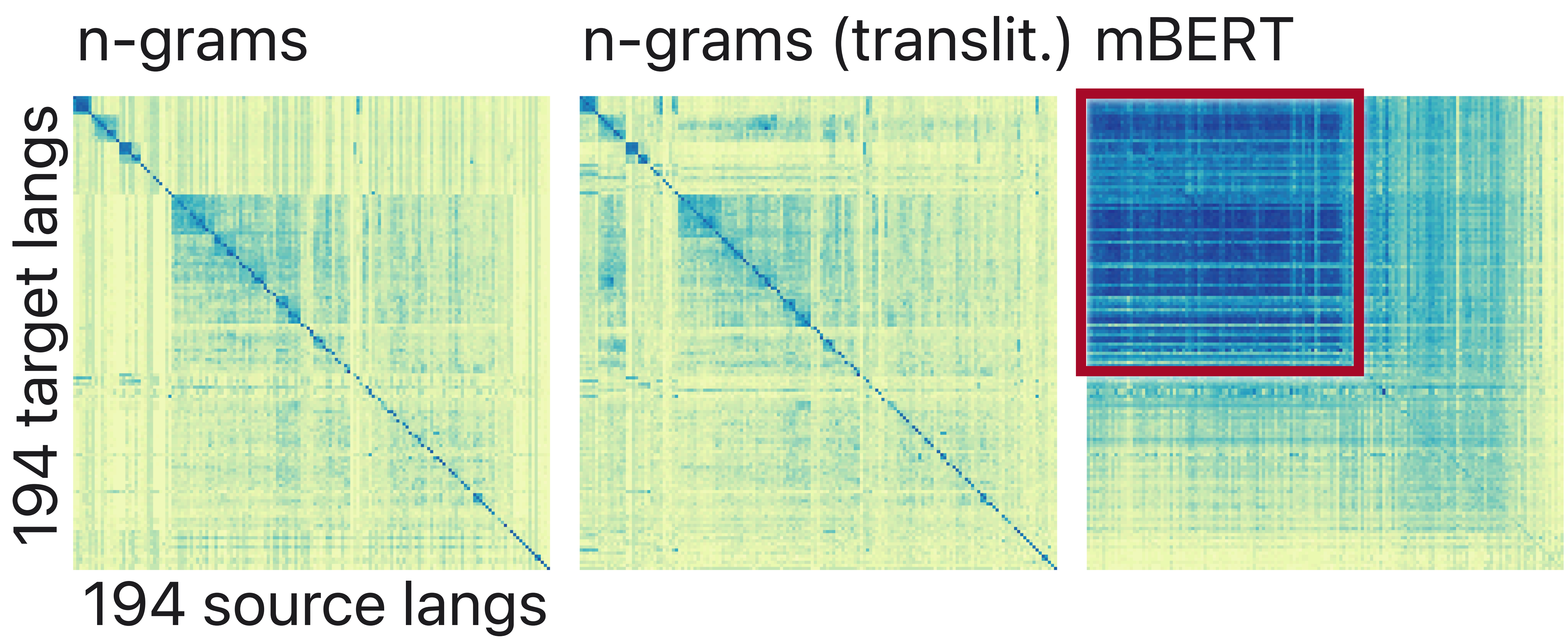
- Universal Dependencies
- 70 × 153 languages
- UDPipe 2 (mono- and multilingual char/word embeddings)

Topic classification

- SIB-200
- 194 × 194 languages
- MLPs → comparable & competitive
- Input representations
  - character n-grams
  - n-grams (transliterated text)
  - mBERT embeddings

SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects (Adelani et al., EACL 2024)

NLP Experiments — topic classification



Topic classification

- SIB-200
- 194 × 194 languages
- MLPs → comparable & competitive
- Input representations
  - character n-grams
  - n-grams (transliterated text)
  - mBERT embeddings

SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects (Adelani et al., EACL 2024)

Overview

Large-scale cross-lingual transfer experiments

Correlations with similarity measures

Practical takeaways for picking source languages

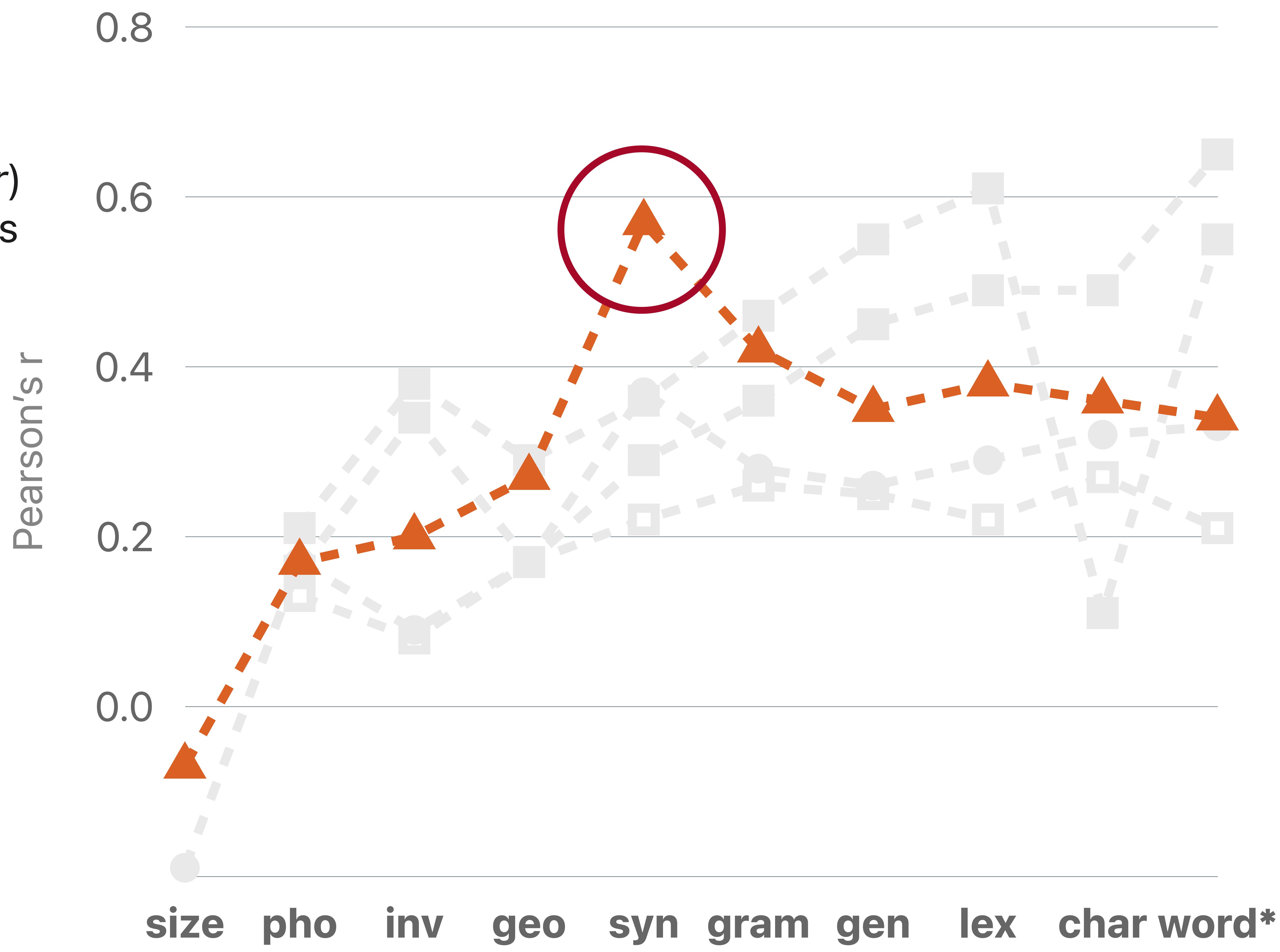
Similarity measures

- Linguistic measures
  - Structural similarities: **Gram**mar, **syn**tax, **pho**nology, phoneme **inv**entory
  - **Lex**ical similarity
  - Phylo**gen**etic relatedness
  - **Geo**graphic proximity
- Dataset measures
  - **Char**acter overlap
  - **Word*** overlap (words, character trigrams, subword tokens)
  - **Size** of training split

Correlations between transfer results and similarity measures

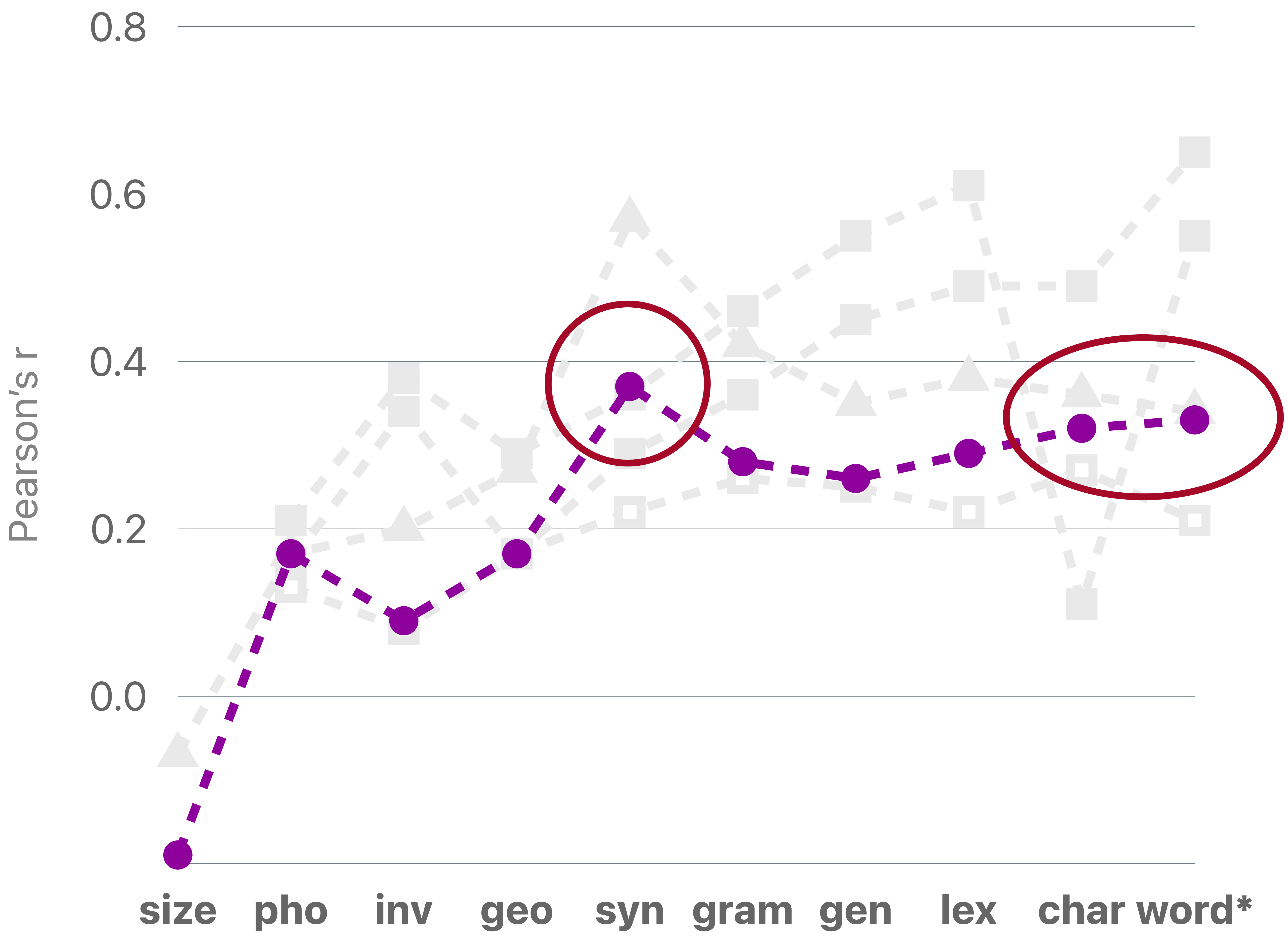
Parsing (labelled attachment score)

Mean correlation (r) over test languages



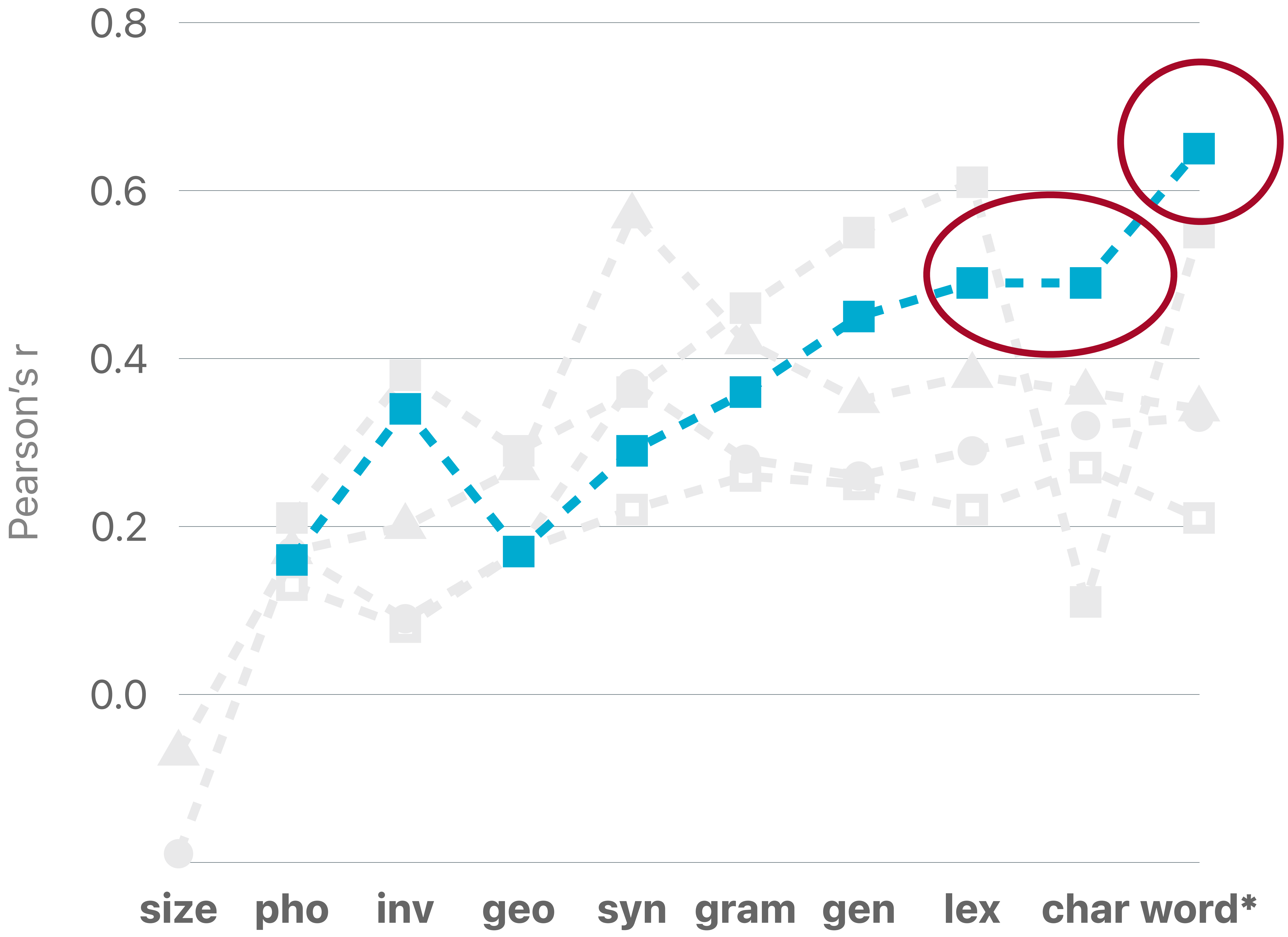
Correlations between transfer results and similarity measures

POS tagging (accuracy)



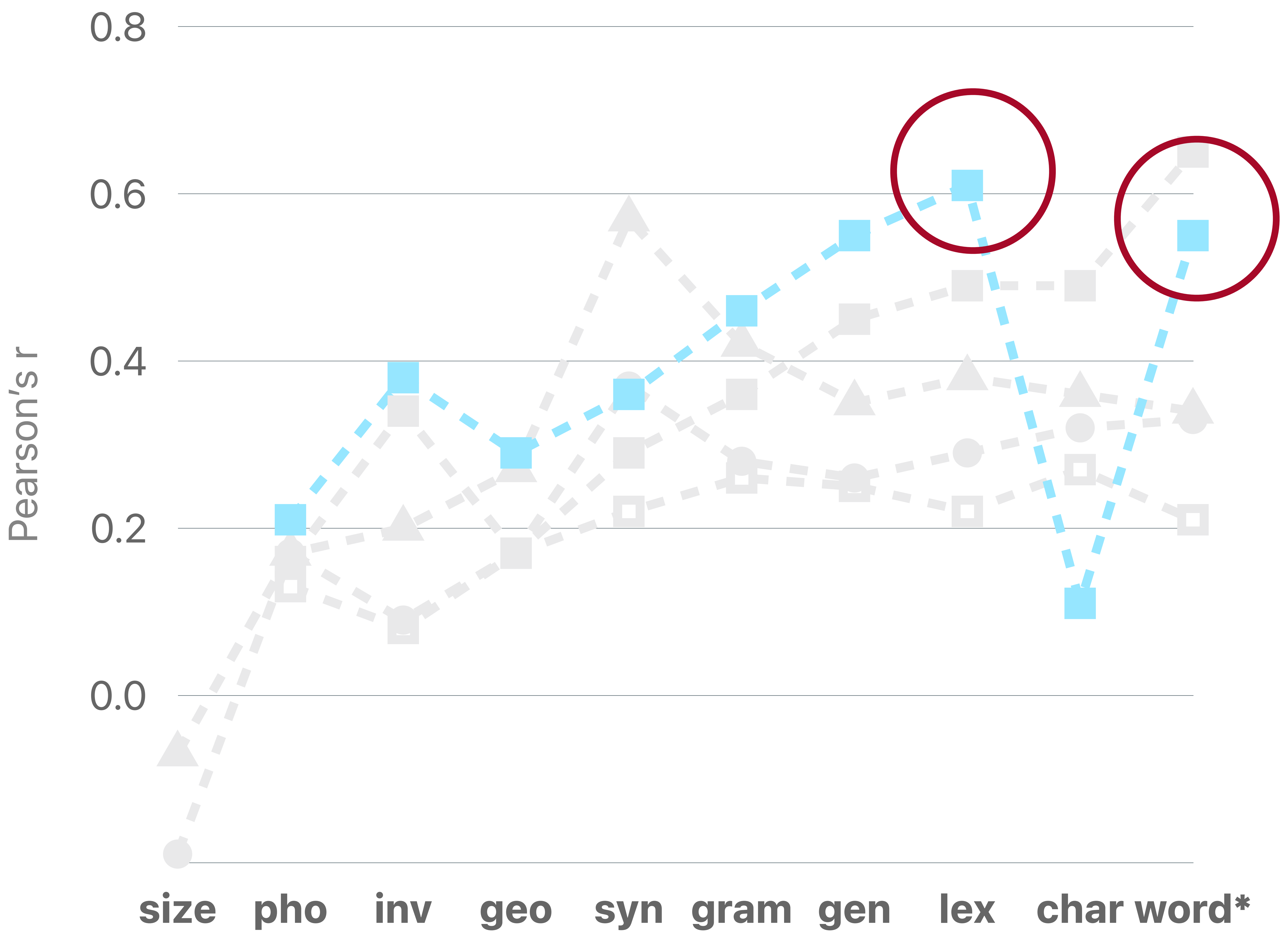
Correlations between transfer results and similarity measures

Topic classification (accuracy) — MLP with n-grams



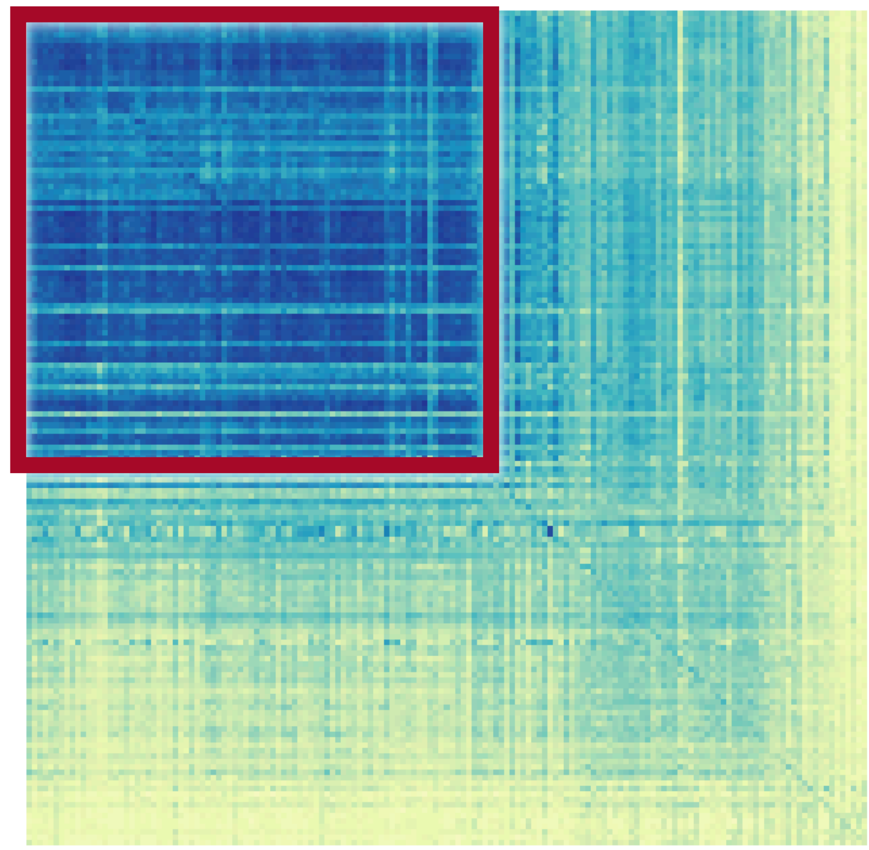
Correlations between transfer results and similarity measures

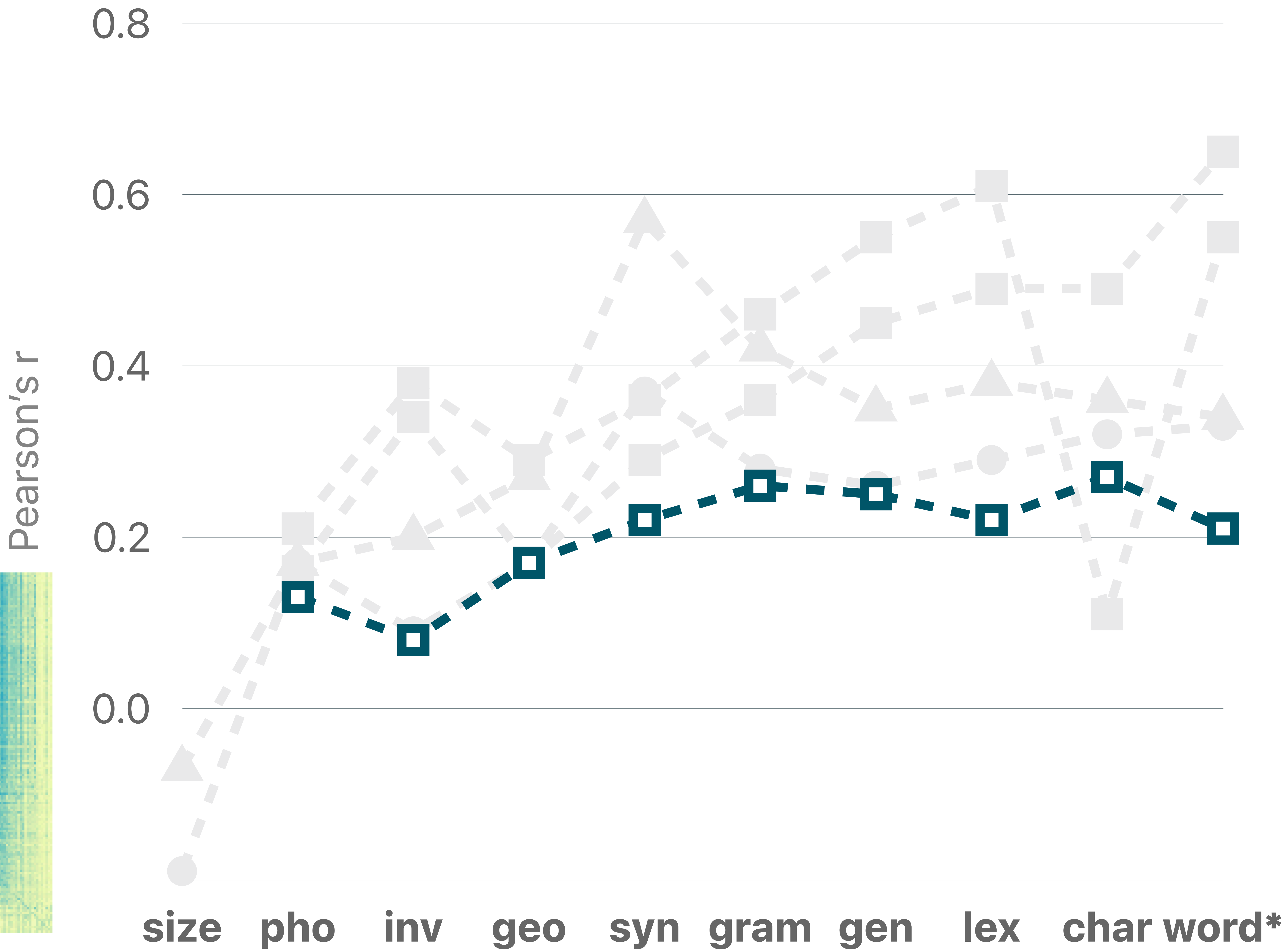
Topic classification (accuracy) — MLP with n-grams (transliterated)



Correlations between transfer results and similarity measures

Topic classification (accuracy) — MLP with mBERT embeddings





Overview

Large-scale cross-lingual transfer experiments

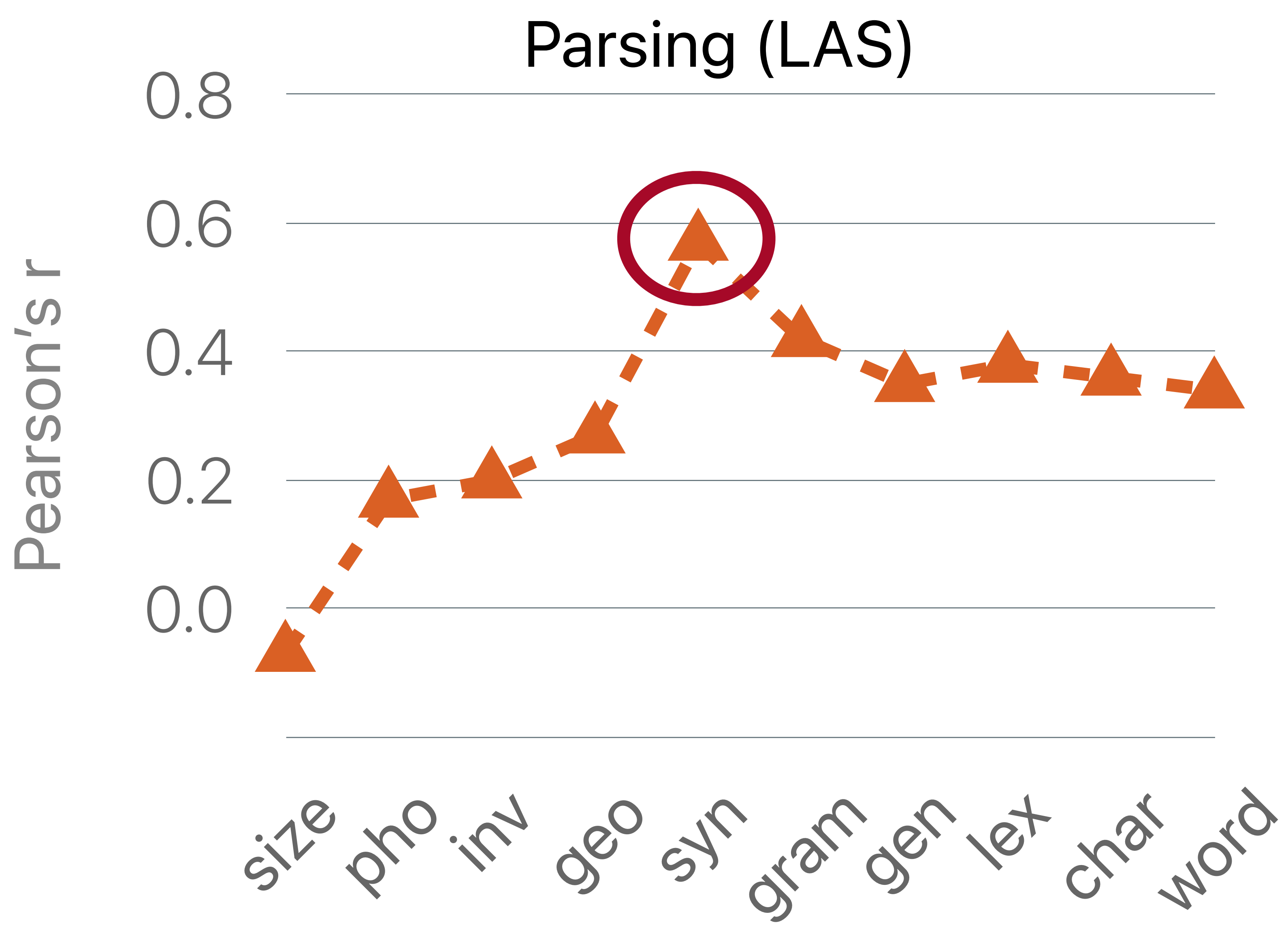
Comparing transfer trends

Practical takeaways for picking source languages

Picking source languages based on similarity measures

	size	pho	inv	geo	syn	gram	gen	lex	char	word*
Top-1 source candidate (= most similar language)										
POS	29	15	14	15	10	12	9	10	15	12
Parsing (LAS)	21	13	13	13	7	10	8	8	16	11
Topics (n-grams)	—	17	17	13	15	14	9	9	13	4
Topics (n-grams, translit.)	—	13	13	11	11	10	7	7	20	3
Topics (mBERT)	—	12	11	10	9	8	8	8	12	9

Mean performance loss in percentage points
if picking the best training language according to one measure
(instead of the overall best one)



Picking source languages based on similarity measures

	size	pho	inv	geo	syn	gram	gen	lex	char	word*
Top-1 source candidate (= most similar language)										
POS	29	15	14	15	10	12	9	10	15	12
Parsing (LAS)	21	13	13	13	7	10	8	8	16	11
Topics (n-grams)	–	17	17	13	15	14	9	9	13	4
Topics (n-grams, translit.)	–	13	13	11	11	10	7	7	20	3
Topics (mBERT)	–	12	11	10	9	8	8	8	12	9
Top-3 source candidates										
POS	25	7	7	5	3	4	5	4	7	5
Parsing (LAS)	18	8	7	5	3	4	4	3	8	6
Topics (n-grams)	–	10	9	5	6	6	5	3	8	2
Topics (n-grams, translit.)	–	8	7	4	5	5	3	3	14	1
Topics (mBERT)	–	6	5	4	4	4	4	4	5	6

Mean performance loss in percentage points
if picking the best training language according to one measure
(instead of the overall best one)

Picking source languages based on other transfer experiments?

	POS	LAS	Topics: n-grams	Topics: n-grams (translit.)	Topics: mBERT
Top-1 source candidate (= best src in another experiment)					
POS	—	2	9	10	16
Parsing (LAS)	1	—	11	10	18
Topics (n-grams)	10	11	—	3	19
Topics (n-grams, translit.)	8	8	3	—	18
Topics (mBERT)	4	5	7	6	—
Top-3 source candidates					
POS	—	1	4	4	7
LAS	0	—	5	5	9
Topics (n-grams)	5	5	—	1	12
Topics (n-grams, translit.)	5	4	2	—	12
Topics (mBERT)	2	2	3	3	—

Mean performance loss in percentage points if picking the best training language according to the results of another transfer experiment (instead of the overall best one)

Conclusions

- Different tasks/input representations
  → different similarity measures matter
- Selecting training languages based on relevant similarity measures (or on similar experiments) works well
  - If possible: compare multiple promising training languages

More details in the paper :)

TM and © 2025 Apple Inc. All rights reserved.