# Language Models Are Polyglots

## Language Similarity Predicts Cross-Lingual Transfer Learning Performance

Juuso Eronen, **Michal Ptaszynski**, Tomasz Wicherkiewicz, Robert Borges, Katarzyna Janic, Zhenzhen Liu, Tanjim Mahmud & Fumito Masui

*Mach. Learn. Knowl. Extr. 2026 | doi:10.3390/make8030065*

# 7,000+ languages.
# Most NLP tools work for ~20.

### High-resource

English, Chinese, etc.
Abundant labeled data

### Low-resource

Most of the world's
languages lack data

### The gap

Cross-lingual transfer
is the bridge

# Everyone just picks English.

But is it always the best source language for transfer?

## Current approach

- Pick English by default
- Or choose source intuitively
- No principled method
- Large performance gaps ignored

## Our approach

- Measure linguistic similarity
- Use it to predict transfer
- Principled source selection
- Better results, less guessing

Core Hypothesis

# The more similar two languages are, the better one transfers to the other.

If true → we can pick optimal source languages systematically, not by pure intuition

METHODOLOGY

# Three independent lines of evidence

We triangulate using causal inference principles

## Similarity Metrics

qWALS, lang2vec, eLinguistics, EzGlot

## Transfer Performance

mBERT & XLM-R
4 tasks, 8 languages

## Expert Survey

3 professional linguists, 769 Qs

BACKGROUND

# The World Atlas of Language Structures

| 2,662 | 192 | ~12% |
| --- | --- | --- |
| Languages | Features | Populated |

## Covers phonology, morphology, syntax & lexicon

- Problem: Most data is missing — English has ~150 features, Danish only 58
- Previous metrics use only a handful of features from WALS (or none at all)
- **Our solution: Use ALL shared features for each language pair separately**

OUR METRIC

# Quantified WALS (qWALS)

A typology-based similarity metric from nearly 200 features

1. Select shared features for each language pair
2. Map values to 0–1 (ordinal encoding)
3. Compute Manhattan distance, normalized
4. Symmetric distance metric for 2,000+ langs

## Key advantages over lang2vec

- Transparent: every feature is inspectable & interpretable
- Task-tunable: features can be re-weighted per downstream task
- Ordinal: preserves gradation (unlike one-hot encoding)

METRICS COMPARED

# Four similarity metrics

| Metric | Focus | Scope | Limitation |
|---|---|---|---|
| eLinguistics | Consonant use | Phonetic only | Accuracy drops for distant languages |
| EzGlot | Lexical overlap | Vocabulary only | Many missing values, asymmetric |
| lang2vec | Multiple sources | Heterogeneous | Opaque weighting, not task-tunable |
| **qWALS (ours)** | ~192 WALS features | Multi-domain | Sparse coverage for some pairs |

**qWALS is the only metric that is both multi-domain AND task-tunable**

# 8 languages, 3 families

| Germanic | Slavic | Koreano-Japonic |
|---|---|---|
| USGB English | PL Polish | JP Japanese |
| DE German | RU Russian | KR Korean |
| DK Danish | HR Croatian | |

*Within-family and cross-family transfer tested in zero-shot setting*

EXPERIMENTAL SETUP

# 4 NLP tasks × 2 models

**DEP** **Dependency Parsing**

Syntax-heavy

**NER** **Named Entity Recognition**

Mixed syntax + semantics

**SA** **Sentiment Analysis**

Semantics-heavy

**ALD** **Abusive Language Identification**

Mixed + pragmatics

## Models

**mBERT** **XLM-R**

= 64 fine-tuned models, each tested on all 8 languages
(4 tasks * 2 models * trained on 8 languages * tested on 8 languages)

# Quick Quiz!

You want to build a cyberbullying detector for Polish.

You have no Polish training data.

Which language should you train on?

**A**

**English**

(most data available)

**B**

**German**

(also Germanic, like English)

**C**

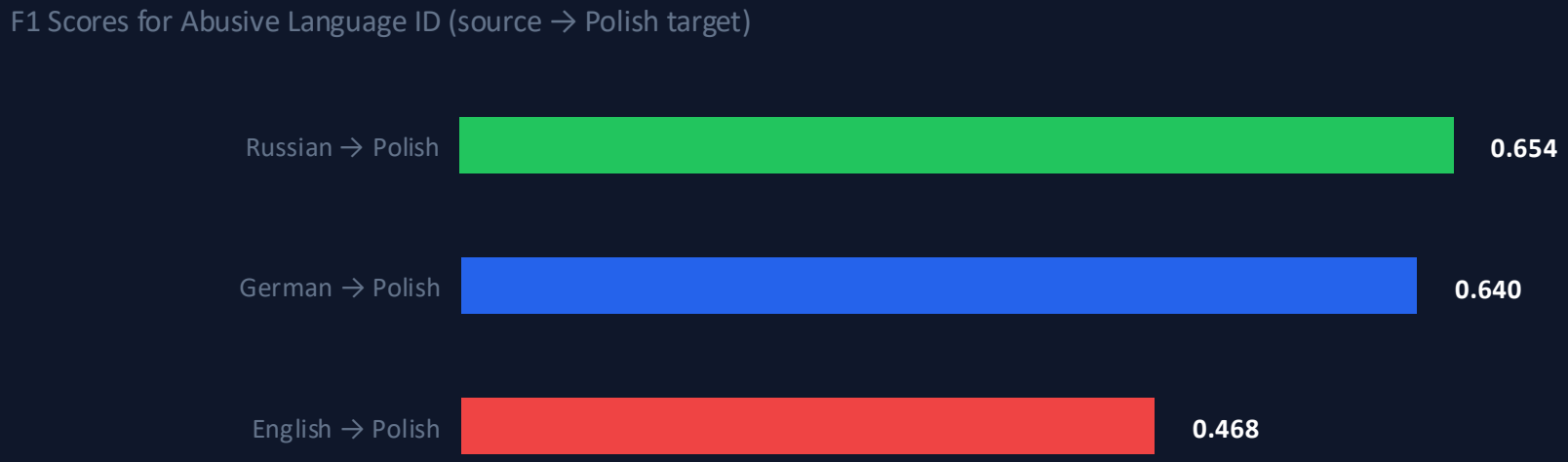
**Russian**

(also Slavic, like Polish)

# Answer: C — Russian

Polish ← Russian transfer: F1 = 0.654 (XLM-R)

Polish ← English transfer: F1 = 0.468



RESULTS

# Transfer performance patterns

### Family matters

Same-family languages transfer best across most tasks

### XLM-R > mBERT

XLM-R outperforms on most tasks, except sentiment analysis

### DEP is hardest

Japanese/Korean → Indo-European: near-zero for dependency parsing

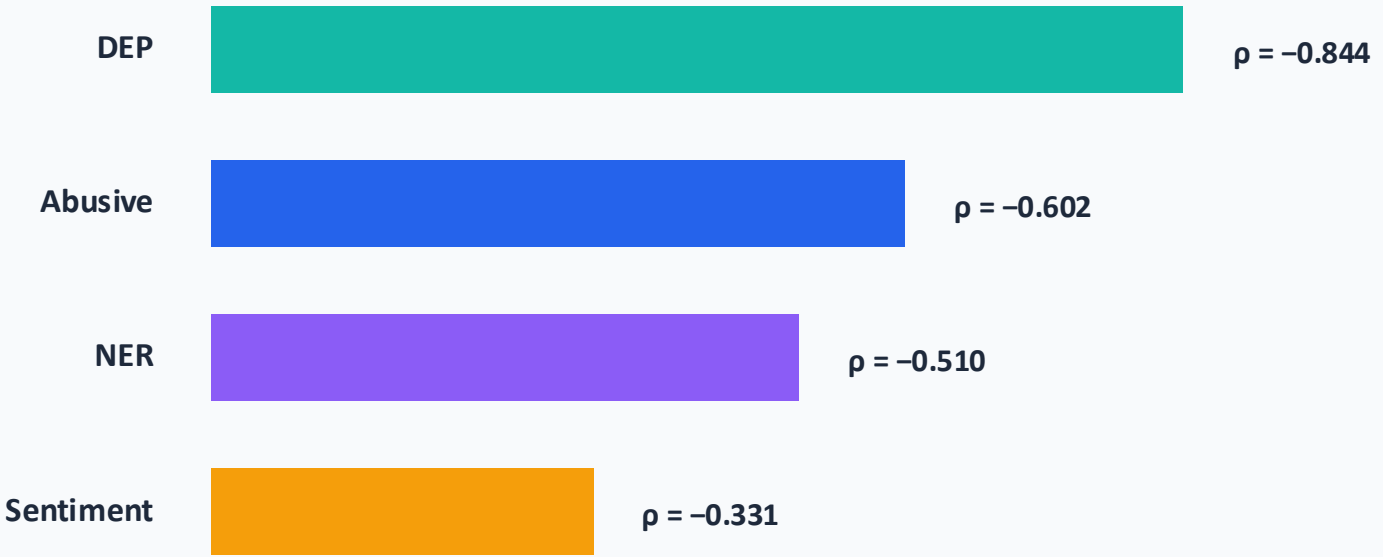
### SA is different

Sentiment analysis shows high transfer even across families

RESULTS

# Similarity strongly correlates with transfer

Spearman's ρ between qWALS (ordinal) and model performance (XLM-R)




*All p-values < 0.001 | Stronger correlation = more syntax-dependent task*

**Dependency parsing correlation with eLinguistics: ρ = −0.897 (near-perfect!)**

# Removing the monolingual "cheat codes"

Same-language pairs inflate correlations. What happens after removing them?

| Task | ρ (all pairs) | ρ (zero-shot only) | Change |
|---|---|---|---|
| **DEP** | −0.844 | **−0.769** | Still strong ✓ |
| **Abusive** | −0.602 | −0.415 | Moderate |
| **NER** | −0.510 | −0.316 | Moderate |
| **Sentiment** | −0.331 | −0.168 | Weak |

DEP stays strong — syntax-driven tasks genuinely depend on linguistic similarity.

Sentiment relies more on semantic/contextual cues that cross family boundaries.

OPTIMIZATION

# Optimizing qWALS per task

Leave-one-feature-out method: drop features that don't help

1. Remove one feature at a time → check correlation change
2. Drop the feature whose removal improves correlation most
3. Repeat until no further improvement

| DEP | |
|---|---|
| Before | −0.771 |
| After | **−0.990** |
| 169 → 75 features | |

| Abusive | |
|---|---|
| Before | −0.646 |
| After | **−0.822** |
| 169 → 53 features | |

| NER | |
|---|---|
| Before | −0.561 |
| After | **−0.808** |
| 169 → 63 features | |

| Sentiment | |
|---|---|
| Before | −0.361 |
| After | **−0.803** |
| 169 → 21 features | |

# −0.990

Pearson correlation for dependency parsing

With just 75 optimized WALS features,
qWALS can near-perfectly predict
which source language will transfer best.

**Syntax-heavy tasks → similarity matters most**

EXPERT SURVEY

# Validating with human experts

**3**
Expert linguists
(20-40 yrs experience)

**769**
Binary-choice
questions

**30**
Languages
covered

**Question format:**

*"Polish is more similar to..."*

**(A) Korean** **(B) Latvian**

Simple binary choice — experts use ALL their knowledge
(grammar, phonology, history, contact, geography)

SURVEY RESULTS

# Experts agree with qWALS ~77% of the time

| Comparison | Agreement | Cohen's κ |
|---|---|---|
| qWALS ↔ Linguist 1 | **76.5%** | 0.755 |
| qWALS ↔ Linguist 2 | **73.0%** | 0.719 |
| qWALS ↔ Linguist 3 | **71.3%** | 0.701 |
| lang2vec ↔ Linguist 1 | **71.9%** | 0.707 |

**qWALS consistently outperforms lang2vec in expert alignment**

## Agreement by task (experts vs. best transfer language):

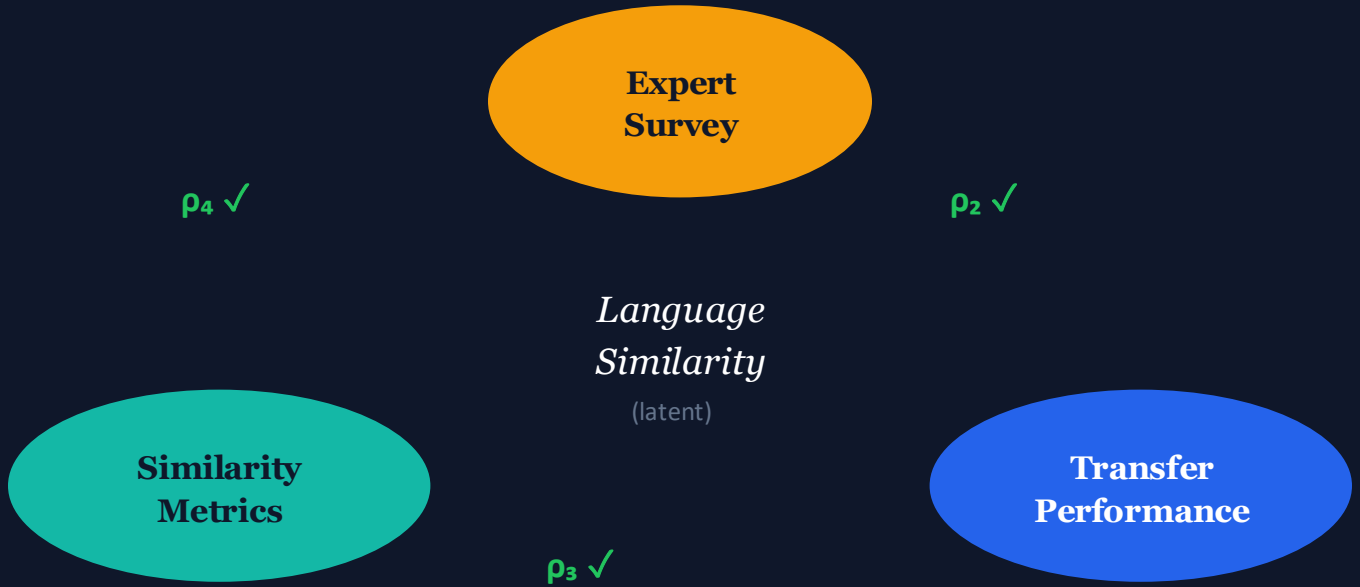
Abusive **100%** agreement

DEP **76.9%** agreement

NER **69.2%** agreement

Sentiment **61.5%** agreement

# Triangulation: All three signals align




*All three pairwise correlations are positive and significant → supports causal role of similarity*

TAKEAWAY

# Don't default to English.

Use linguistic similarity to pick a smarter source language.

- qWALS: open-source, transparent, task-tunable similarity metric
- Covers 2,662 languages — including many low-resource ones
- Optimized feature sets per task improve prediction dramatically
- Expert validation confirms computational metrics match human judgment
- A similar source language usually outperforms English

**qWALS is released as open source — use it for your own research!**

# Honest caveats & next steps

## Limitations

- 8 languages, 3 families — need broader coverage (follow-up paper already in review)
- WALS sparsity: some pairs share <30% features
- No Romance languages tested yet
- Only encoder models (mBERT, XLM-R)
- Correlation ≠ Causation (despite triangulation)

## Future directions

- Add Romance cluster (FR, ES, PT, IT)
- Integrate Grambank + PHOIBLE data
- Test on decoder-only LLMs (GPT-4, LLaMA)
- Use KANs for non-linear feature learning
- Continuous pretraining & probing studies
- Add online DEMO

# Thank You!

Questions, comments, … DEMO!

https://pypi.org/project/qwals/

doi.org/10.3390/make8030065

michal@mail.kitami-it.ac.jp | jeronen@mail.kitami-it.ac.jp

**qWALS is open source — try it on your language pair!**