

Automated part-of-speech tagging across 52 language varieties

Benchmarking multilingual NLP for scalable clinical language analysis

Loretta Gasparini ^{1,2,3}, Daniel Shaw ², Thayabaran Kathiresan ^{1,2}, Devon Borysiewicz ², Geoffrey W. Stuart ^{1,2}, Cross-linguistic POS Consortium, Adam P. Vogel ^{1,2,3}

¹The University of Melbourne, Australia | ²Redenlab Ltd | ³Murdoch Children's Research Institute, Australia

105

linguists

58

language varieties

52

with NLP POS taggers

42

countries

14

language families

18

writing systems

Background

Language endpoints are meaningful clinical outcomes in neurological and neurodevelopmental conditions. Digital language outcomes can be derived from conversations, monologues or picture descriptions using natural language processing (NLP). POS tagging quantifies grammatical word classes, but model accuracy across languages is often reported with limited transparency.

Objective

Develop and report a standard method for evaluating POS taggers across languages, then provide proof-of-concept results using a short, standardized passage.

Methods

- 1 North Wind & Sun passage (113 words in English)
- 2 Manual POS tags by trained linguists
- 3 Consensus ground truth where ≥ 2 linguists
- 4 4 reproducible NLP libraries Stanza, spaCy, FreeLing, CLASSLA
- 5 Token realignment then POS agreement
- 6 MCC + 95% CIs plus Cohen's κ for IRR

Caveat

This is a proof-of-concept benchmark, not a clinical deployment validation.

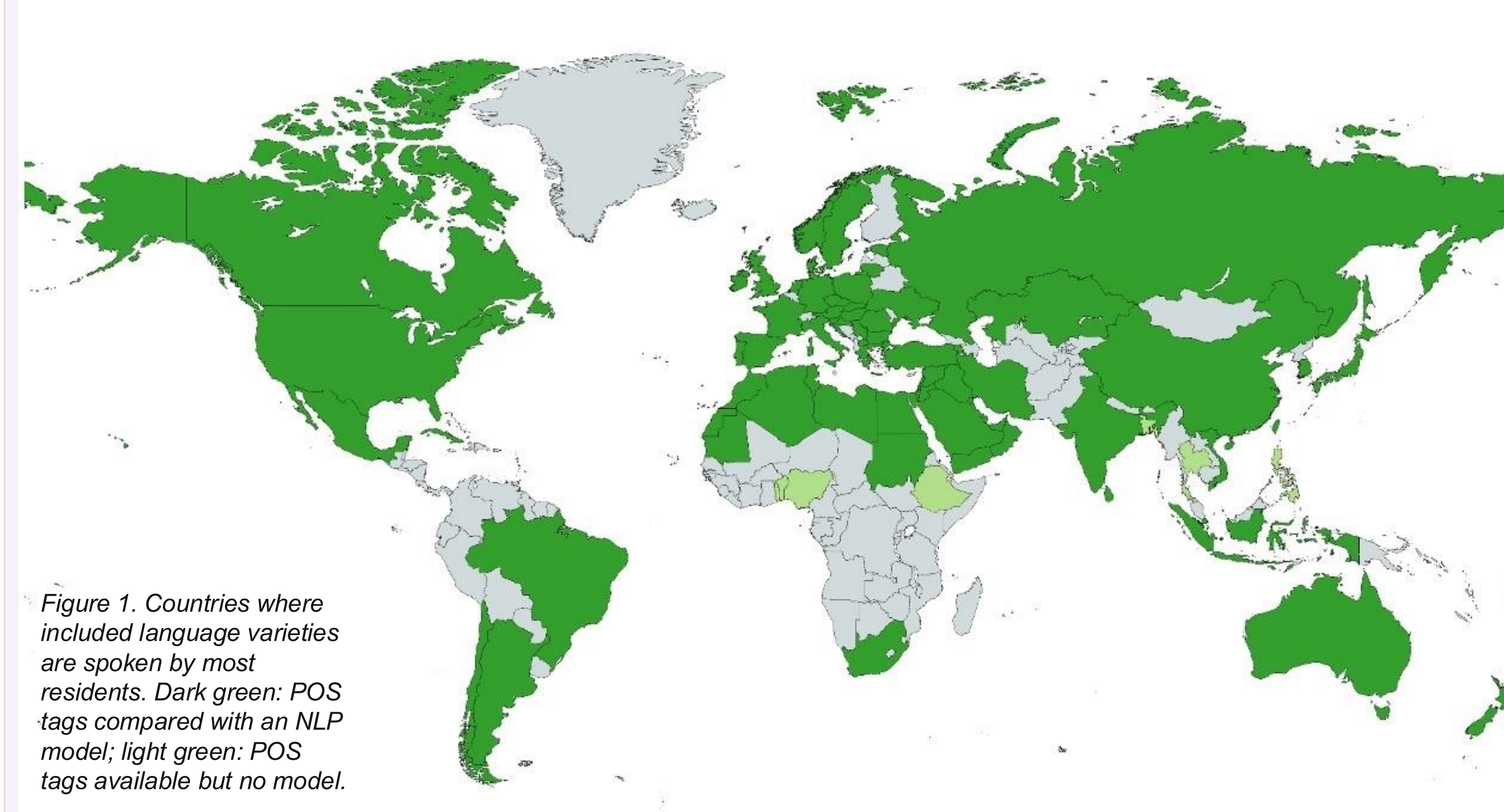
The passage was short and written. We recommend replication using longer, ecologically valid language samples before applying outputs to clinical research decisions. Tentative precision target: many uses may require approximately 1,000–2,500 words to narrow MCC confidence intervals. Decision context matters: individual clinical decisions need higher accuracy than group-level research comparisons.

Conclusion

This benchmark shows that multilingual POS tagging is feasible but uneven. Twenty-five language varieties have promising NLP performance under a transparent methodology, while 27 require additional ground-truth work, tokenization improvement, or model development before clinical language endpoints can be trusted at scale.

Clinical endpoint implication: NLP-derived language features should be validated by language variety, stimulus type and intended clinical decision context, not assumed from English or library-reported accuracy.

Global language coverage



Primary result: language-variety triage

52 language varieties with NLP models

25 Strong 11 Unclear 16 Poor

Strong: $>80\%$ linguist IRR, $>95\%$ token match, $>90\%$ NLP–linguist agreement.

Unclear: did not meet one or two thresholds.

Poor: did not meet any threshold or had critical reliability/tokenization/POS-agreement problems.

Performance profile

31 / 44
IRR comparisons had $\kappa \geq 80\%$
where individual linguist tags were available

7
tokenization failures
Arabic, Trad. Chinese, Farsi, Japanese, Marathi, Turkish, Vietnamese

3
highest-certainty models
French, Italian, Slovenian had MCC CI $\geq 90\%$

Strong, unclear and poor groups

Strong / promising

Armenian, Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, Estonian, French, Galician, Greek, Hebrew, Italian, Lithuanian, Norwegian, Polish, Romanian, Russian, Slovenian, Spanish (Argentina, Chile, Cuba, Spain), Swedish

Unclear

Asturian, English, German, Hindi, Hungarian, Serbian, Slovak, Spanish (Mexico), Portuguese (Brazil, Portugal), Ukrainian

Poor

Afrikaans, Arabic, Chinese (Simplified, Traditional), Farsi, Indonesian, Japanese, Kazakh, Korean, Macedonian, Marathi, Tamil, Telugu, Turkish, Vietnamese, Welsh

What limits accuracy?

Tokenization errors

Low token match rate can remove tokens from POS agreement and distort downstream estimates.

Impacted: 7 varieties

POS disagreement

Some models showed lower agreement with linguists than library documentation would suggest.

Key cases: Afrikaans, Simplified Chinese, Korean, Macedonian

Interpretation and next steps

Replicate

Promising languages need longer, ecologically valid samples before clinical use.

Improve

Poor-performing languages need better labelled data, tokenization and model fine-tuning.

Clarify

Unclear cases need stronger ground truth and/or more reliable linguist agreement.

Further details

Scan the QR code for author details, data, code, figures and references.

