A Reproducibility Study on Quantifying Language Similarity: The Impact of Missing Values in the URIEL Knowledge Base

Hasti Toossi[†], Guo Qing Huai[†], Jinyu Liu[†],
Eric Khiu[*], A. Seza Doğruöz[#], En-Shiun Annie Lee[†,‡]
[†]University of Toronto, Canada [*]University of Michigan, USA
[#]LT3, IDLab, Ghent University, Belgium [‡]Ontario Tech University, Canada
hasti.toossi@mail.utoronto.ca as.dogruoz@ugent.be annie.Lee@ontariotechu.ca

Motivation

Understanding Language Similarity

- Essential for multilingual NLP applications.
- No consensus on best methods for low-resource languages (LRLs).

Challenges with URIEL Database

- URIEL (Littell et. al., 2017) is a typological knowledge base that aggregates linguistic information from various sources
- Inconsistencies of definitions and missing values affect reliability of language similarity measurements.

Our Contributions

- Identify areas for improvement for URIEL and its lang2vec tool for calculating language similarity.
- Analyze feature coverage across languages
- Investigated how URIEL is utilized in current NLP research, highlighting the dependency on its accuracy.

Literature Review

Usage of Language Similarity

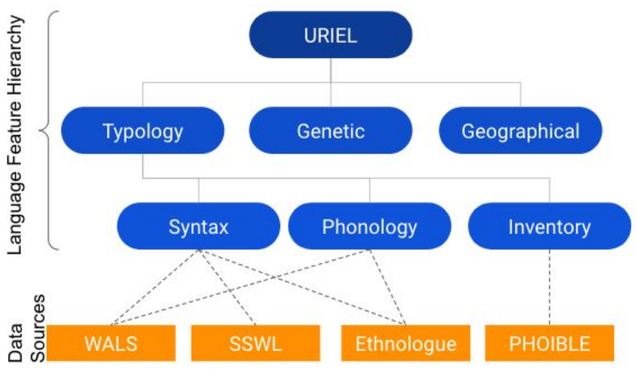
- **Cross-lingual modelling and learning** (Lauscher et al., 2020)
- **Performance prediction** (Patankar et al., 2022, Xia et al., 2020, Srinivasan et al., 2021)
- **Cross-lingual transfer and language translation** (Lin et al., 2019, Huang et al., 2021)
- **Integration with language models** (Üstün et al., 2020, Adilazuarda et al., 2024)

Observed Limitation

- Predicted values exhibit noticeable clusters due to biases introduced by family-based prediction of missing values (Ponti et al., 2019).

URIEL/lang2vec's Method

1 Collect information from various sources



Contains missing values!

Fig. 1: URIEL feature hierarchy and data sources

2 Take an aggregate of the different sources

- Union
- Average
- *k*-nearest neighbor (KNN)

Unclear which one to use in which scenario!

*k*NN details not provided

3 Compute the distance of feature vectors of the two languages

- **lang2vec documentation:** Cosine distance

$$D_C(u, v) := 1 - S_C(u, v)$$

- **URIEL documentation:** Angular distance

$$D_\theta(u, v) := \frac{1}{\pi} \arccos(S_C(u, v))$$

where $S_C(u, v) := \frac{u \cdot v}{\|u\| \|v\|}$

Conflicting definition

Maybe regularize?

Reproducibility

Handling Missing Values

By inspection, the following approach is most likely used by URIEL to handle missing values:

- **All values missing:** Replace entire vector with [1, 1, …, 1].
- **Partial values missing:** Replace missing entries with 0.

Percentage of Reproducible Distances

Aggregate Vector	Distance Metric	syntactic	phonological	inventory
union	cosine	23.90%	61.62%	40.04%
	reg_angular	**93.96%**	**95.42%**	**99.45%**
average	cosine	23.95%	61.62%	40.04%
	reg_angular	89.82%	95.21%	90.53%
knn	cosine	0.39%	1.45%	0.12%
	reg_angular	2.46%	2.53%	9.70%

- Union vector with regularized angular distance achieves the highest reproducibility of pre-calculated distances.
- Union and average vectors are identical for many languages ⇒ Similar reproducibility.
- Some distance values cannot be reproduced using any method- unclear factor of irreproducibility.

Feature Coverage

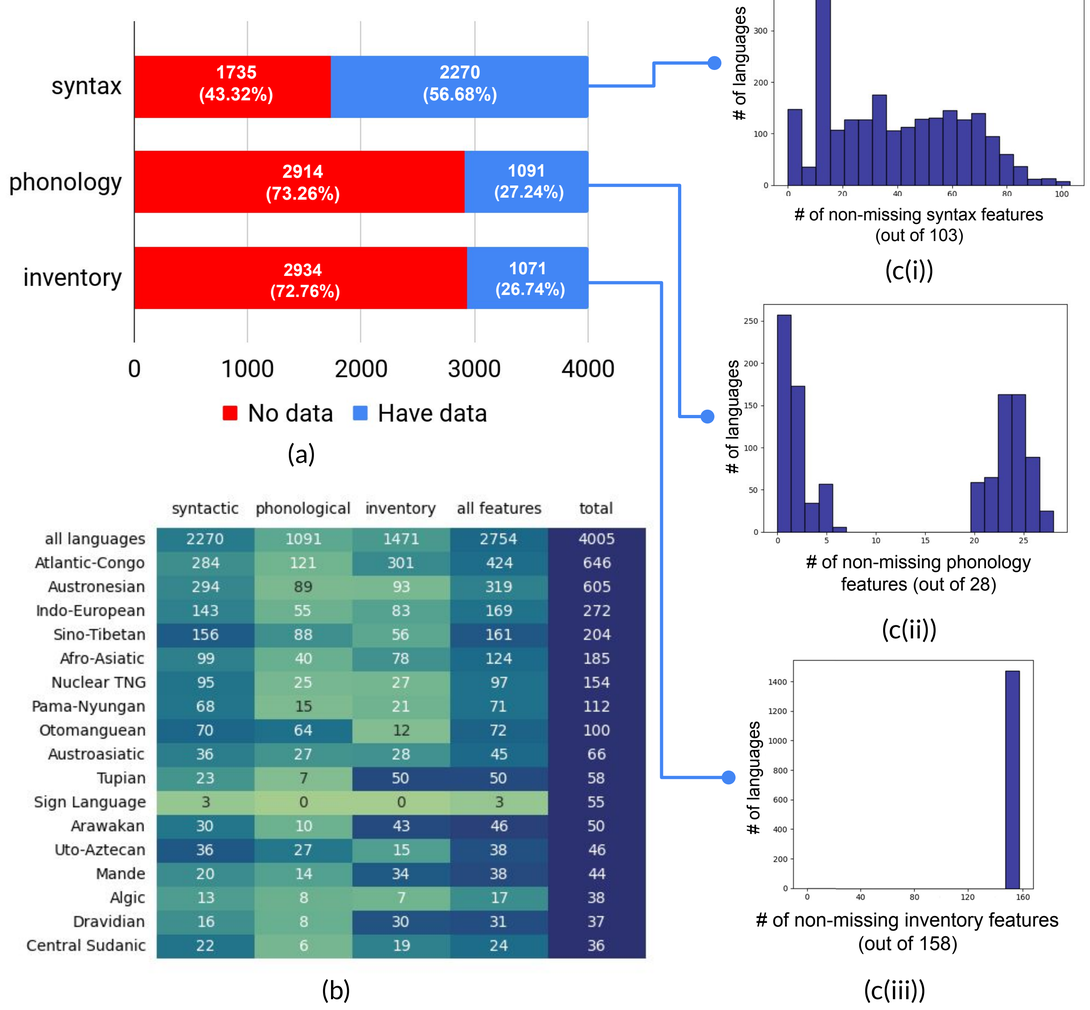


Fig. 2: Feature coverage across 4005 languages (overall: a; by family: b) and distribution based on number of non-missing features in union vector for syntax_union (c(i)), phonology_union (c(ii)), and inventory_union (c(iii)), excluding languages with empty feature vectors.

Conclusion & Next Step

- **Unclear Definitions:** Definitions of distance values in the documentation is unclear; some values remain irreproducible.
- **Missing Values:** Approaches to handle missing values has no clear justification- affects validity of language distances involving LRLs.
- **Low Coverage:** 31.24% of languages lack feature information, making provided distances meaningless.
- **Future Directions:**
  - Establish clear guidelines for acceptable levels of missing data.
  - Explore alternative similarity measurement for LRLs.

Acknowledgement

We thank Pratik Nadipelli for his valuable assistance in the literature review, Aditya Khan, Phuong Hanh Hoang, and Mason Shipton for their meticulous efforts in proofreading the paper.

References

- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain. Association for Computational Linguistics.