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, Universiteit Gent, Belgium [‡] Ontario Tech University, Canada
hasti.toossi@mail.utoronto.ca as.dogruoz@ugent.be annie.lee@cs.toronto.edu

Abstract

In the pursuit of supporting more languages around the world, tools that characterize properties of languages play a key role in expanding the existing multilingual NLP research. In this study, we focus on a widely used typological knowledge base, URIEL, which aggregates linguistic information into numeric vectors. Specifically, we delve into the soundness and reproducibility of the approach taken by URIEL in quantifying language similarity. Our analysis reveals URIEL's ambiguity in calculating language distances and in handling missing values. Moreover, we find that URIEL does not provide any information about typological features for 31% of the languages it represents, undermining the reliability of the database, particularly on low-resource languages. Our literature review suggests URIEL and lang2vec are used in papers on diverse NLP tasks, which motivates us to rigorously verify the database as the effectiveness of these works depends on the reliability of the information the tool provides.

1 Introduction

Categorizing and quantifying variations and similarities between languages is critical for applications such as building multilingual large language models (Xia et al., 2020; Nllb team, 2022), examining the effects of cross-lingual transfer (Lin et al., 2019b), understanding code-switching between languages (Doğruöz et al., 2021; Doğruöz and Sitaram, 2022), selecting pivot languages when translating from one language to another (Wu and Wang, 2007), or sharing language tools (Strassel and Tracey, 2016). However, there is no consensus on how to measure the similarity between languages due to the difficulty and subjectivity involved in assessing various aspects of languages. This challenge becomes even more pronounced when dealing with low-resource languages (Joshi et al., 2020), where limited linguistic knowledge is available to researchers.

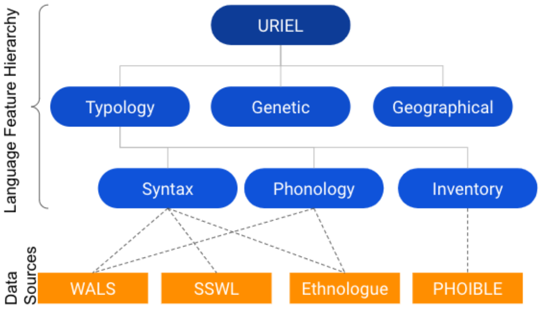


Figure 1: URIEL Feature Hierarchy and Data Sources.

URIEL is a knowledge base that aggregates linguistic information for 4,005 languages from various data sources (Figure 1) and computes distances based on this information. The lang2vec tool provides an interface for querying URIEL (Littell et al., 2017). In many of the 198 citations of URIEL and lang2vec, the distance values and feature vectors provided by URIEL have been used to quantify language similarity and categorize language features.

In this study, we analyze the URIEL database to assess its capacity as a resource for quantifying language similarity. We evaluate the reproducibility and validity of the methodology employed in calculating language similarity measurements. We also examine the language and feature coverage of URIEL, which affects the meaningfulness of the vectors and distance values.

In addition, we conducted a literature review of papers that cite URIEL to gain a better understanding of the influence of URIEL in these works.

2 Methodology for Reproducibility

2.1 Description of URIEL

For a pair of languages, URIEL computes the distance of the corresponding features through the following steps:

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 233–241
June 18, 2024 ©2024 Association for Computational Linguistics

1. Collect information from various sources for a specific feature.

2. Take an aggregate of the different sources for a single feature.

3. Compute the distance of the feature vectors of the two languages.

URIEL knowledge base URIEL unifies information from various sources (Figure 1), such as WALS (Dryer and Haspelmath, 2013), SSWL (Koopman, 2009), PHOIBLE (Moran and McCloy, 2019), Ethnologue (Eberhard and Fennig, 2023), and Glottolog (Hammarström et al., 2023). The *features* of a language are broken down into three types:

1. Typological features `syntax`, `phonology` and `inventory`, which describe the corresponding linguistic characteristics of the language.

2. Phylogenetic feature `family`, which specifies the language families to which the language belongs.

3. Geographical feature `geography` for the approximate location where the language is most commonly spoken in the world.

All features are described using binary (0 or 1) vectors to represent language facts. Missing values are marked by "`--`".

For each feature, different vectors are provided depending on the source. For instance, URIEL provides `syntax` vectors sourced from each of WALS, SSWL, and Ethnologue. Similarly, other feature vectors are derived from multiple sources.

Aggregating sources Since the information for each feature can be taken from several sources, URIEL uses three aggregation methods to consolidate feature information: union, average, or k-nearest neighbours (kNN).

For the union aggregation, denoted using the union operator "`|`", each feature is set to 1 if any of the sources for that feature has a value of 1. If the feature value is 0 in all sources, the feature is set to 0. If the feature value is missing in all sources, the union result has a missing entry, denoted by "`--`".

For the average aggregation, each entry is the mean across all sources in which it appears. This result is a value between 0 and 1, with a non-binary value if there are disagreements among the sources. The feature is missing, denoted "`--`", if the value is missing in all sources.

Lastly, for the kNN aggregation, the missing values are predicted based on languages similar in terms of genetic, geographic and featural distances. It is unclear how aggregation is done for kNN, as the details are omitted from the URIEL paper. Littell et al. (2017) writes: "We will describe these procedures, the exact notions of distance involved, alternative prediction methods that we also investigated, and their results in more detail in a future article."

Computing language distances For each language pair, URIEL provides pre-calculated distance values based on the aggregated feature vectors. While the exact methodology for distance calculations is not specified in the URIEL paper (Littell et al., 2017), additional documentation for URIEL and lang2vec provides two different distance calculation methods.

The lang2vec documentation[1] uses cosine distance to compute distances between feature vectors. The cosine distance D_C between two vectors u and v is defined as

$$D_C(u, v) := 1 - S_C(u, v) \tag{1}$$

where S_C is cosine similarity defined by

$$S_C(u, v) := \frac{u \cdot v}{\|u\|\|v\|}. \tag{2}$$

On the other hand, the URIEL documentation[2] defines a distance equivalent to angular distance. The angular distance D_θ between two vectors u and v is defined as

$$D_\theta(u, v) := \frac{1}{\pi} \arccos(S_C(u, v)) \tag{3}$$

where S_C is the same cosine similarity defined in (2).

Note that the value of $D_\theta(u, v)$ can range between 0 and 0.5 because all feature values are positive. However, distances in URIEL range between 0 and 1, with "0 representing identity and 1 being as far apart as two languages can be" based on the URIEL documentation. Therefore, it is reasonable to assume that this distance metric is regularized to $2D_\theta(u, v)$.

[1] lang2vec is the Python tool developed by the authors for querying URIEL. https://github.com/antonisa/lang2vec

[2] http://www.cs.cmu.edu/~dmortens/projects/7_project/

Aggregate	Distance	All Languages			Languages with Non-Empty Feature Vectors		
Vector	Metric	syntactic	phonological	inventory	syntactic	phonological	inventory
union	cosine	23.90%	61.62%	40.04%	14.24%	34.28%	0.07%
union	angular	**93.36%**	**95.42%**	**99.45%**	**95.89%**	**87.76%**	**98.52%**
average	cosine	23.95%	61.62%	40.04%	14.38%	34.28%	0.07%
average	angular	89.82%	95.21%	90.53%	88.92%	86.90%	71.23%
knn	cosine	0.39%	1.45%	0.12%	0.39%	1.45%	0.12%
knn	angular	2.46%	2.53%	9.70%	2.46%	2.53%	9.70%

Table 1: Percentage of all language pairs with reproducible distances (up to 2 decimal places) using each method.

3 Results

3.1 Reproducibility Study

We attempted to reproduce the pre-calculated distance provided by URIEL for each language pair. This involved reproducing the aggregation step and the distance calculation step using the feature vectors provided by URIEL. We used the aggregated feature vectors from URIEL as the basis for the distance computations.

Reproducing aggregated vectors While we successfully reproduced the first two aggregation vectors (union and average), we were unable to replicate the exact kNN aggregation vector because the necessary details were not provided.

Reproducing distance calculations As mentioned earlier, URIEL provides pre-computed distances for each language pair based on different feature vectors (particularly syntactic, phonological, and inventory). However, the methodology used to calculate these distances is unclear in the documentation. We aimed to reproduce these provided distance values to infer the methodology used.

There are three ambiguities in the documentation regarding distance computations:

1. Which aggregated vector is used; union, average, or knn?

2. Which distance metric is used; cosine distance $D_C(u, v)$ or regularized angular distance $2D_\theta(u, v)$?

3. How are the missing feature values treated?

We found that among possible methods of treating missed values, the following method aligns with the pre-computed distances closely:

- If every value in a feature vector is missing, replace it with a vector of the same length containing only 1's.
- If some, but not all, values in a vector are missing, replace the missing values with 0.

Using this method for treating missing values, we calculated the distances for each language pair using all possible combinations of aggregation methods and distance metrics.

The percentage of all language pairs whose distance can be reproduced with each set of choices is shown in Table 1 ("All Languages" section)[3]. The highest percentage of reproducible distances was achieved using regularized angular distance with union vectors.

A similarly high percentage of distances could be reproduced by using regularized angular distance with average vectors instead of union vectors. This can be explained by noting that the union and average vectors are identical for many languages. Corresponding average and union vectors are equal when all available sources agree on the relevant features of a language. Specifically, syntax_union and syntax_average are equal for 95.23% of languages, phonology_union and phonology_average for 99.73% of languages, and inventory_union and inventory_average for 91.59% of languages.

Additionally, in Table 1 ("Languages with Non-Empty Feature Vectors" section), we consider the reproducibility of distances for language pairs where both languages have non-empty feature vectors. This is relevant because all empty vectors are considered identical for distance purposes.

We conclude that regularized angular distance with union vectors is the most likely method used to calculate the pre-computed distance vectors provided by URIEL. However, some distance values could not be reproduced using this or any other method we tried. There are no clear factors causing the irreproduciblity of certain distance values.

[3] URIEL provides phonological and inventory distances up to 4 decimal points. However, reproducibility suffers when using more than 2 decimal points.

3.2 Analysis of Feature Coverage

URIEL provides feature vectors for 4,005 languages, as well as corresponding distance values for all pairs of these languages (16,040,025 pairs).

However, a large number of features in these vectors have missing values, and many feature vectors are completely empty, indicating that every feature in these vectors is missing. This raises concerns about the meaningfulness of the distance values provided for such languages, as the vectors contain no information to distinguish these languages.

Out of the 4,005 languages, 1,735 (43.32%) have empty `syntax_union` vectors, 2,914 (72.76%) have empty `phonology_union` vectors, and 2,534 (63.27%) have empty `inventory_union` vectors. Furthermore, 1,251 (31.24%) languages have no feature information at all, meaning they have empty vectors for `syntax_union`, `phonology_union`, and `inventory_union`. Some languages have empty vectors in one or more of these three categories, but not all.

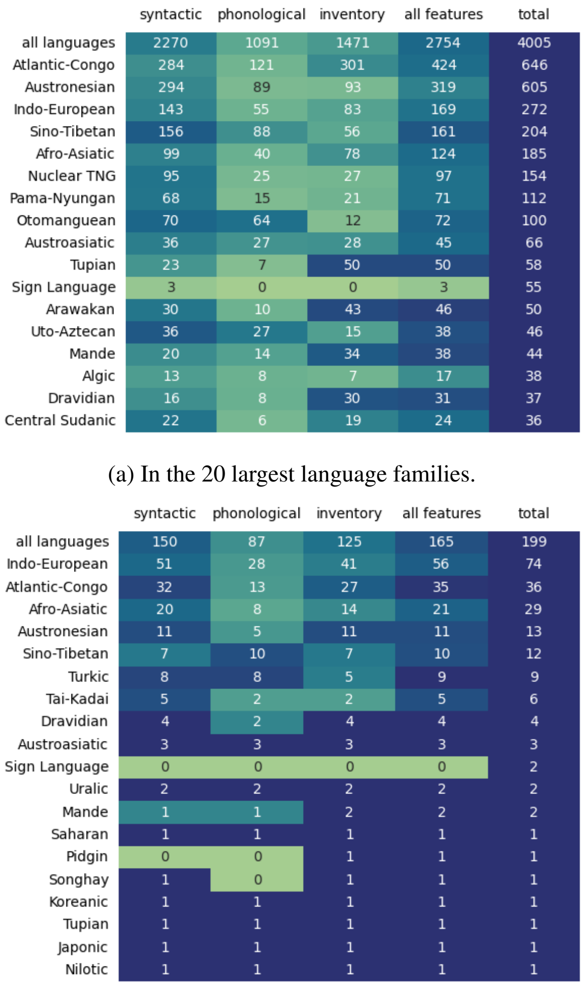
Figure 2a shows the number of languages with non-empty `union` feature vectors in each of the 20 largest language families. The column labelled "all features" represents the number of languages with non-empty `union` feature vectors in at least one of the categories. The "total" column shows the total number of languages from each language family included in URIEL. The shading indicates the percentage of languages with non-empty vectors compared to the total number of languages in the corresponding language family.

Similarly, Figure 2b focuses on the top 200 most spoken languages in the world[4], as identified by Ethnologue 2023. Figure 5 presents this information for all language families in URIEL.

3.3 Distribution of Non-Missing Features

In section 3.2, we discussed languages with empty feature vectors, i.e., languages that lack any feature information in a given category. We found that these languages constitute a large portion of all languages in the URIEL dataset.

To better understand the feature coverage of the remaining languages, we will now exclude the languages with empty feature vectors. In Figure 3, we visualize the distribution of the remaining languages based on the number of feature values provided in their `union` vectors for each category.



	syntactic	phonological	inventory	all features	total
all languages	2270	1091	1471	2754	4005
Atlantic-Congo	284	121	301	424	646
Austronesian	294	89	93	319	605
Indo-European	143	55	83	169	272
Sino-Tibetan	156	88	56	161	204
Afro-Asiatic	99	40	78	124	185
Nuclear TNG	95	25	27	97	154
Pama-Nyungan	68	15	21	71	112
Otomanguean	70	64	12	72	100
Austroasiatic	36	27	28	45	66
Tupian	23	7	50	50	58
Sign Language	3	0	0	3	55
Arawakan	30	10	43	46	50
Uto-Aztecan	36	27	15	38	46
Mande	20	14	34	38	44
Algic	13	8	7	17	38
Dravidian	16	8	30	31	37
Central Sudanic	22	6	19	24	36

(a) In the 20 largest language families.

	syntactic	phonological	inventory	all features	total
all languages	150	87	125	165	199
Indo-European	51	28	41	56	74
Atlantic-Congo	32	13	27	35	36
Afro-Asiatic	20	8	14	21	29
Austronesian	11	5	11	11	13
Sino-Tibetan	7	10	7	10	12
Turkic	8	8	5	9	9
Tai-Kadai	5	2	2	5	6
Dravidian	4	2	4	4	4
Austroasiatic	3	3	3	3	3
Sign Language	0	0	0	0	2
Uralic	2	2	2	2	2
Mande	1	1	2	2	2
Saharan	1	1	1	1	1
Pidgin	0	0	1	1	1
Songhay	1	0	1	1	1
Koreanic	1	1	1	1	1
Tupian	1	1	1	1	1
Japonic	1	1	1	1	1
Nilotic	1	1	1	1	1

(b) In the top 200 most spoken languages.

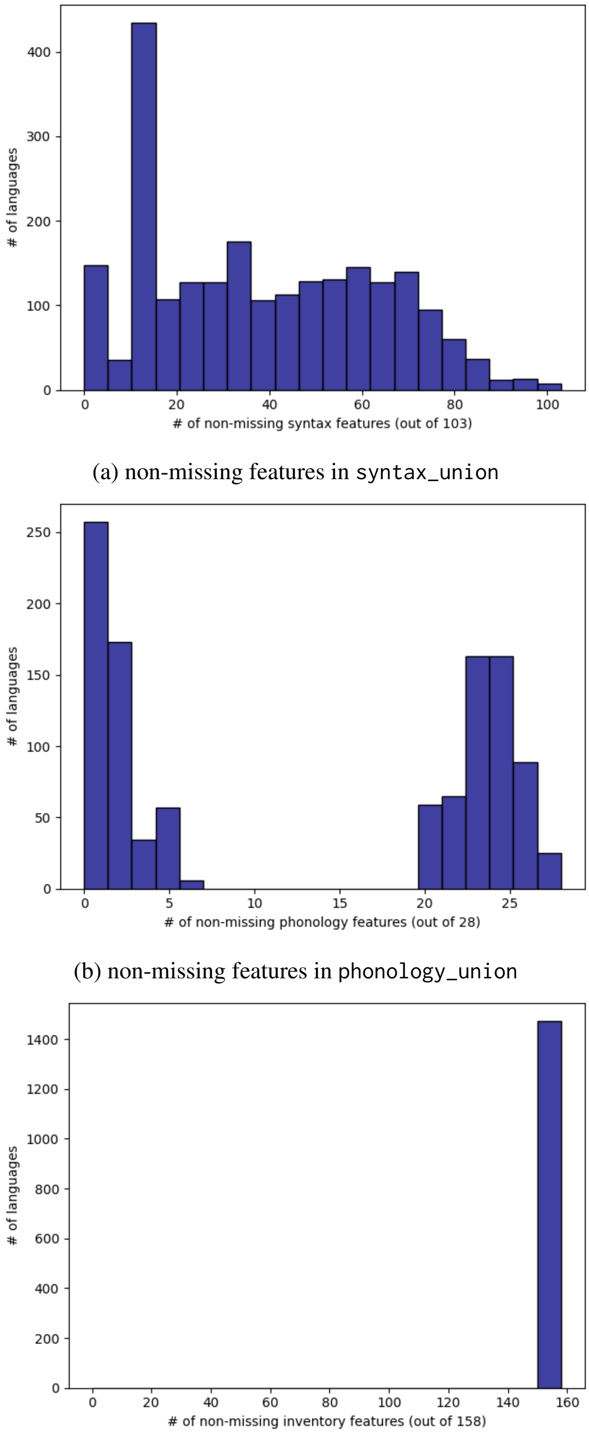
Figure 2: Number of languages with non-empty `union` feature vectors

In Figure 3c, we observe that if any language has a non-empty `inventory_union` vector, then this vector contains no missing values. By referencing the original source[5], we find that this source provides complete International Phonetic Alphabet (IPA) charts for all languages it covers. Since `inventory` vectors represent the information from the IPA chart of each language, they do not have any missing values when a complete IPA chart is available.

As depicted in Figure 3b,languages with non-empty `phonology_union` vectors generally have either at least 20 or at most 7 phonology features, with no values in between. On the other hand, syntax features exhibit a more even distribution (Figure 3a) with a peak in the number of languages with 11 to 15 syntax features.

[4] Excluding Bajjika, the 103rd most spoken language, which is missing from URIEL.

[5] In this case, the relevant source is PHOIBLE, a repository of cross-linguistic phonological inventory data.



(a) non-missing features in `syntax_union`

(b) non-missing features in `phonology_union`

(c) non-missing features in `inventory_union`

Figure 3: Distribution of languages based on the number of non-missing features in the `union` vector for each category, excluding languages with empty feature vectors.

4 Literature Review

4.1 Methodology of Literature Review

A structured search strategy was implemented to gather articles containing citations of URIEL/lang2vec from Google Scholar, sorted by relevance. We then reviewed each paper through a particular process. First, we read the abstract and introduction of the paper to fill in the summary section, and identified relevant keywords from each paper. Then, we used the search function to find occurrences of "Littel", "URIEL", and "lang2vec" to locate where and how URIEL was used in the paper. Finally, we searched for keywords such as "database", "language distance", and "WALS" to identify other methods co-existing with or compared to URIEL in the paper.

Following the initial search, duplicated instances of URIEL usage and articles with similar topics were categorized. Further analysis focused on the most cited articles, as well as articles relevant to performance prediction, language distance, and typological feature comparison. Selected articles underwent a full-text review, during which a detailed examination of methodologies, findings and limitations was conducted.

4.2 Findings from Literature Review

Our literature review consists of a comprehensive analysis of 198 citations of the URIEL database up to February 2024. The cited literature focuses on a range of topics, including cross-lingual modelling, performance prediction, and other NLP applications such as document image classification, text-to-speech, and speech recognition (Adams et al., 2019; Raj et al., 2023).

Researchers have explored the efficacy of URIEL in cross-lingual modelling, cross-lingual learning, and zero-shot transfer scenarios (Lauscher et al., 2020). Patankar et al. (2022), Xia et al. (2020), and Srinivasan et al. (2021) delved into methodologies for predicting the performance of multilingual NLP models across diverse tasks.

Researchers often use URIEL and lang2vec to select the source language in cross-lingual transfer tasks and language translation tasks. Lin et al. (2019a) attempt to solve the task of automatically selecting optimal transfer languages as a ranking problem and build models that consider URIEL's language features to perform this prediction. Huang et al. (2021) use lang2vec to verify model outcomes, evaluate the effectiveness of models across different languages, and analyze the correlation between model outcomes and language distance between the source and target languages in language translation tasks. Aside from language distance computation, Üstün et al. (2020) integrated lang2vec into models such as BERT and multilayer perceptrons, enhancing their performance across various linguistic tasks.

Adilazuarda et al. (2024) demonstrated a way to align lang2vec feature vectors and Multilingual BERT (mBERT) embeddings to explore whether multilingual language models (MLMs), such as mBERT, capture the linguistic constraints defined by URIEL vectors. Based upon theobservation that mBERT embeddings and lang2vec vectors strongly correlate, the paper introduces a new method(LINGUALCHEMY) that aligns model representations with the linguistic knowledge by leveraging URIEL vectors. This is achieved by adding an additional URIEL loss term to the regular classification loss. URIEL loss is defined as the mean squared error (MSE) between projected model output and the corresponding URIEL vectors.

Notably, Ponti et al. (2019) highlighted the issue of predicted World Atlas of Language Structures (WALS) values from URIEL exhibiting noticeable clusters, due to biases introduced by family-based prediction of missing values in URIEL.

5 Conclusion

In conclusion, in our attempt to reproduce URIEL's "language distances", we identified several areas for improvement:

- **Unclear definitions:** The documentation for the definition of distance values provided by URIEL is unclear. Through our attempts, we identified the likely definitions used, but there are some distance values that remain irreproducible for unknown reasons.

- **Missing Values:** When computing distances, missing values in the feature vectors are handled by replacement with 0. There is no clear justification for this approach, which affects distance values for languages with many missing values (e.g., with a majority being the low-resource languages).

- **Low Coverage:** We found that 31.24% of the languages in URIEL have no linguistic feature information. While language distance values are provided for these languages by URIEL, they are not meaningful due to the empty feature vectors. While the low coverage leads to a broader issue for low-resource languages, which is difficult to solve, URIEL can address this by providing a nan value in these cases, which would make it clearer to the user when language distance values cannot be meaningfully derived.

As demonstrated in our literature review, there are broad use cases for measuring language similarity. By understanding and addressing areas of improvement for URIEL and lang2vec, we can contribute to the progress of research in multilingualism and language diversity, especially for low-resource languages that are not properly represented by these knowledge bases and tools.

5.1 Future Work

For future research, we are planning to establish clear guidelines for acceptable levels of missing data in linguistic datasets. Secondly, we aim to refine URIEL specifically for medium- and high-resource languages. For low-resource languages, we will explore alternative similarity measurements, such as conceptual distance or other overlooked linguistic features. Our objective is to advance computational linguistics research by tackling missing data challenges and improving method applicability across diverse linguistic contexts.

5.2 Limitations

The limitation of this research is its reliance on the accuracy and completeness of the URIEL knowledge base when extracting data from its sources. Any inaccuracies or omissions within the URIEL dataset could impact the reproducibility and reliability of our findings. We did not verify the reliability of the external data sources, nor did we compare them against URIEL.

Acknowledgement

We thank Pratik Nadipelli for his valuable assistance in the literature review, Aditya Khan, Fiona Hoang, and Mason Shipton for their meticulous efforts in proofreading the paper. Lastly, we appreciate our NAACL24 mentor William Held for his valuable feedback on our first draft.

References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Alham Fikri Aji, Genta Indra Winata, and Ayu Purwarianti. 2024. Lingualchemy: Fusing typological and geographical elements for unseen language generalization.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Gary F. Simons Eberhard, David M. and Charles D. Fennig. 2023. Ethnologue: Languages of the world.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8.

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Hilda Koopman. 2009. Syntactic structures of the world's languages (sswl). *Hemendik hartua: http://SSWL. railsplayground. net/(azken bisita 2017-03-05)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019a. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019b. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

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*, volume 2, pages 8–14.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

James Cross Onur cCelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Alison Youngblood Bapi Akula Loïc Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon L. Spruit C. Tran Pierre Yves rews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzm'an Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang Nllb team, Marta Ruiz Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. To train or not to train: Predicting the performance of massively multilingual models. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 8–12, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Anjali Raj, Shikhar Bharadwaj, Sriram Ganapathy, Min Ma, and Shikhar Vashishth. 2023. Masr: Multi-label aware speech representation.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21:165–181.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646.

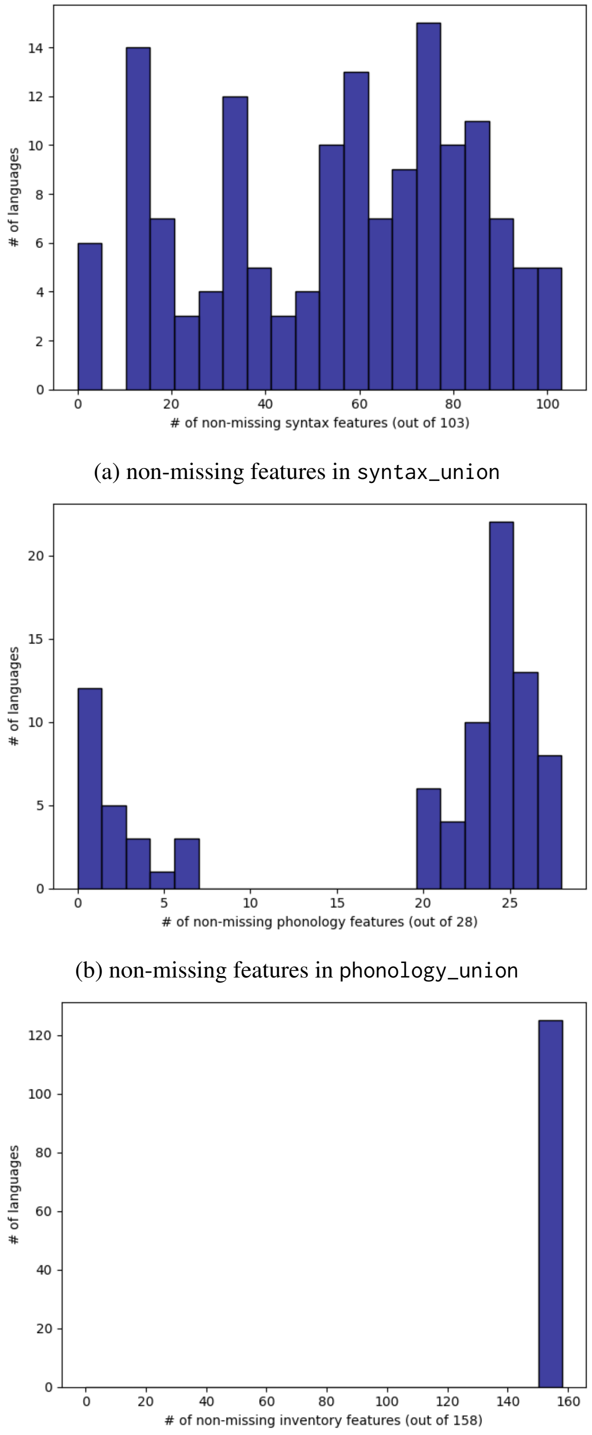
A Top 200 Most Spoken Languages

Figure 4 provides information similar to Figure 3 in the main text, but focuses on the top 200 most spoken languages in the world, as identified by Ethnologue 2023, instead of all 4,005 languages in URIEL.

Note that Bajjika, the 103rd most spoken language in the world (with 12.3M speakers), is missing from URIEL. Consequently, figures 2b and 4 include data only for the other 199 languages.

B Full Table of Coverage Based on Language Family

Figure 5 shows the number of languages with non-empty feature vectors for each language family in URIEL. Figure 2a in the main text is an abridged version that displays only the 200 largest language families.

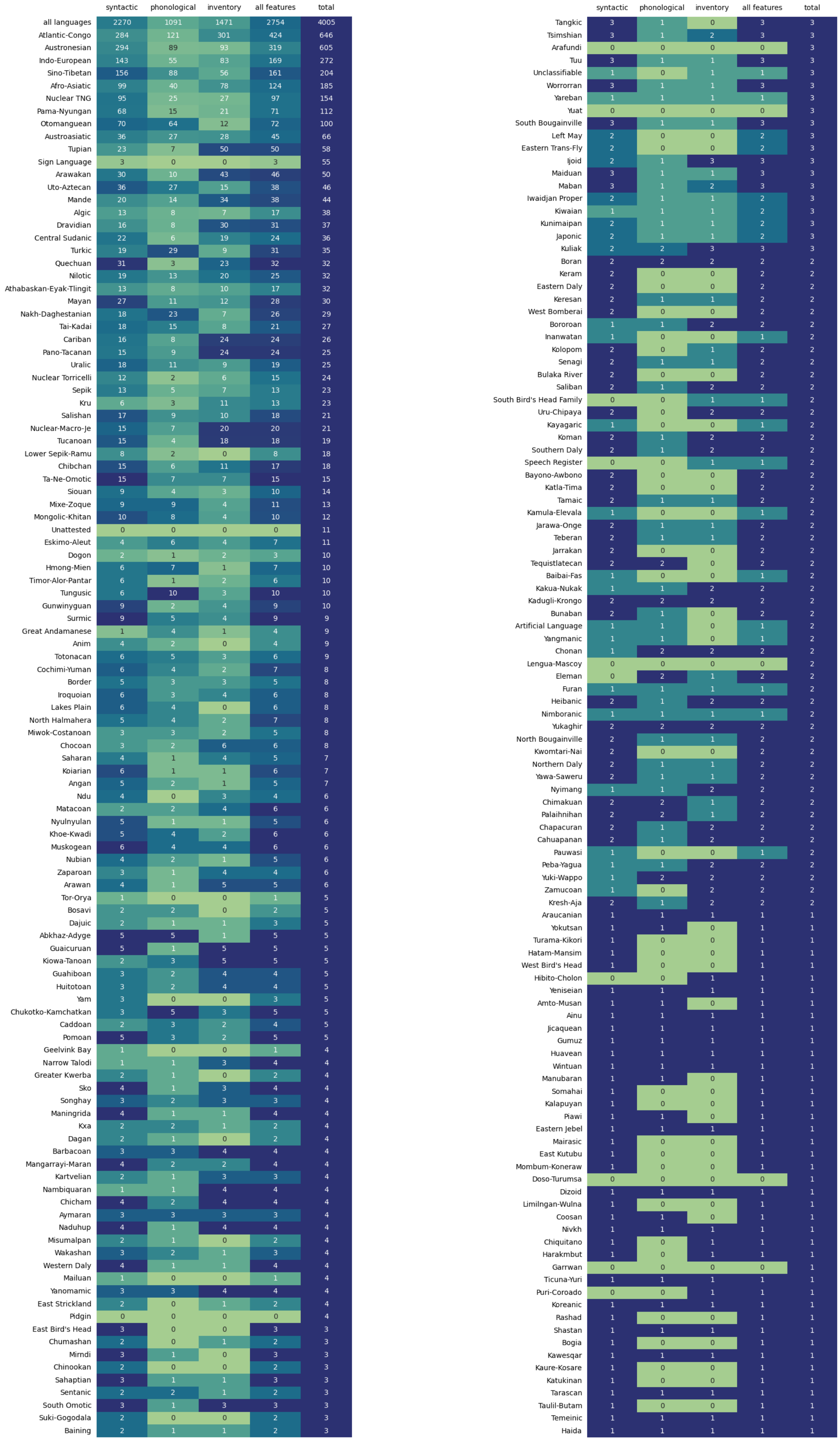


(a) non-missing features in `syntax_union`

(b) non-missing features in `phonology_union`

(c) non-missing features in `inventory_union`

Figure 4: Distribution of the top 200 most spoken languages based on the number of non-missing features in the `union` vector for each category, excluding languages with empty feature vectors.



	syntactic	phonological	inventory	all features	total
all languages	2270	1091	1471	2754	4005
Atlantic-Congo	284	121	301	424	646
Austronesian	294	89	93	319	605
Indo-European	143	55	83	169	272
Sino-Tibetan	156	88	56	161	204
Afro-Asiatic	99	40	78	124	185
Nuclear TNG	95	25	27	97	154
Pama-Nyungan	68	15	21	71	112
Otomanguean	70	64	12	72	100
Austroasiatic	36	27	28	45	66
Tupian	23	7	50	50	58
Sign Language	3	0	0	3	55
Arawakan	30	10	43	46	50
Uto-Aztecan	36	27	15	38	46
Mande	20	14	34	38	44
Algic	13	8	7	17	38
Dravidian	16	8	30	31	37
Central Sudanic	22	6	19	24	36
Turkic	19	29	9	31	35
Quechuan	31	3	23	32	32
Nilotic	19	13	20	25	32
Athabaskan-Eyak-Tlingit	13	8	10	17	32
Mayan	27	11	12	28	30
Nakh-Daghestanian	18	23	7	26	29
Tai-Kadai	18	15	8	21	27
Cariban	16	8	24	24	26
Pano-Tacanan	15	9	24	24	25
Uralic	18	11	9	19	25
Nuclear Torricelli	12	2	6	15	24
Sepik	13	5	7	13	23
Kru	6	3	11	13	23
Salishan	17	9	10	18	21
Nuclear-Macro-Je	15	7	20	20	21
Tucanoan	15	4	18	18	19
Lower Sepik-Ramu	8	2	0	8	18
Chibchan	15	6	11	17	18
Ta-Ne-Omotic	15	7	7	15	15
Siouan	9	4	3	10	14
Mixe-Zoque	9	9	4	11	13
Mongolic-Khitan	10	8	4	10	12
Unattested	0	0	0	0	11
Eskimo-Aleut	4	6	4	7	11
Dogon	2	1	2	3	10
Hmong-Mien	6	7	1	7	10
Timor-Alor-Pantar	6	1	2	6	10
Tungusic	6	10	3	10	10
Gunwinyguan	9	2	4	9	10
Surmic	9	5	4	9	9
Great Andamanese	1	4	1	4	9
Anim	4	2	0	4	9
Totonacan	6	5	3	6	9
Cochimi-Yuman	6	4	2	7	8
Border	5	3	3	5	8
Iroquoian	6	3	4	6	8
Lakes Plain	6	4	0	6	8
North Halmahera	5	4	2	7	8
Miwok-Costanoan	3	3	2	5	8
Chocoan	3	2	6	6	8
Saharan	4	1	4	5	7
Koiarian	6	1	1	6	7
Angan	5	2	1	5	7
Ndu	4	0	3	4	6
Matacoan	2	2	4	6	6
Nyulnyulan	5	1	1	5	6
Khoe-Kwadi	5	4	2	6	6
Muskogean	6	4	4	6	6
Nubian	4	2	1	5	6
Zaparoan	3	1	4	4	6
Arawan	4	1	5	5	6
Tor-Orya	1	0	0	1	5
Bosavi	2	2	0	2	5
Dajuic	2	1	1	3	5
Abkhaz-Adyge	5	5	1	5	5
Guaicuruan	5	1	5	5	5
Kiowa-Tanoan	2	3	5	5	5
Guahiboan	3	2	4	4	5
Huitotoan	3	2	4	4	5
Yam	3	0	0	3	5
Chukotko-Kamchatkan	3	5	3	5	5
Caddoan	2	3	2	4	5
Pomoan	5	3	2	5	5
Geelvink Bay	1	0	0	1	4
Narrow Talodi	1	1	3	4	4
Greater Kwerba	2	1	0	2	4
Sko	4	1	3	4	4
Songhay	3	2	3	3	4
Maningrida	4	1	1	4	4
Kxa	2	2	1	2	4
Dagan	2	1	0	2	4
Barbacoan	3	3	4	4	4
Mangarrayi-Maran	4	2	2	4	4
Kartvelian	2	1	3	3	4
Nambiquaran	1	1	4	4	4
Chicham	4	2	4	4	4
Aymaran	3	3	3	3	4
Naduhup	4	1	4	4	4
Misumalpan	2	1	0	2	4
Wakashan	3	2	1	3	4
Western Daly	4	1	1	4	4
Mailuan	1	0	0	1	4
Yanomamic	3	3	4	4	4
East Strickland	2	0	1	2	4
Pidgin	0	0	0	0	4
East Bird's Head	3	0	0	3	3
Chumashan	2	0	1	2	3
Mirndi	3	1	0	3	3
Chinookan	2	0	0	2	3
Sahaptian	3	1	1	3	3
Sentanic	2	2	1	2	3
South Omotic	3	1	3	3	3
Suki-Gogodala	2	0	0	2	3
Baining	2	1	1	2	3
Tangkic	3	1	0	3	3
Tsimshian	3	1	2	3	3
Arafundi	0	0	0	0	3
Tuu	3	1	1	3	3
Unclassifiable	1	0	1	1	3
Worrorran	3	1	1	3	3
Yareban	1	1	1	1	3
Yuat	0	0	0	0	3
South Bougainville	3	1	1	3	3
Left May	2	0	0	2	3
Eastern Trans-Fly	2	0	0	2	3
Ijoid	2	1	3	3	3
Maiduan	3	1	1	3	3
Maban	3	1	2	3	3
Iwaidjan Proper	2	1	1	2	3
Kiwaian	1	1	1	2	3
Kunimaipan	2	1	1	2	3
Japonic	2	1	1	2	3
Kuliak	2	2	3	3	3
Boran	2	2	2	2	2
Keram	2	0	0	2	2
Eastern Daly	2	0	0	2	2
Keresan	2	1	1	2	2
West Bomberai	2	0	0	2	2
Bororoan	1	1	2	2	2
Inanwatan	1	0	0	1	2
Kolopom	2	0	1	2	2
Senagi	2	1	1	2	2
Bulaka River	2	0	0	2	2
Saliban	2	1	2	2	2
South Bird's Head Family	0	0	1	1	2
Uru-Chipaya	2	0	2	2	2
Kayagaric	1	0	0	1	2
Koman	2	1	2	2	2
Southern Daly	2	1	2	2	2
Speech Register	0	0	1	1	2
Bayono-Awbono	2	0	0	2	2
Katla-Tima	2	0	0	2	2
Tamaic	2	1	1	2	2
Kamula-Elevala	1	0	0	1	2
Jarawa-Onge	2	1	1	2	2
Teberan	2	1	1	2	2
Jarrakan	2	0	0	2	2
Tequistlatecan	2	2	0	2	2
Baibai-Fas	1	0	0	1	2
Kakua-Nukak	1	1	2	2	2
Kadugli-Krongo	2	2	2	2	2
Bunaban	2	1	0	2	2
Artificial Language	1	1	0	1	2
Yangmanic	1	1	0	1	2
Chonan	1	2	2	2	2
Lengua-Mascoy	0	0	0	0	2
Eleman	0	2	1	2	2
Furan	1	1	1	1	2
Heibanic	2	1	2	2	2
Nimboranic	1	1	1	1	2
Yukaghir	2	2	2	2	2
North Bougainville	2	1	1	2	2
Kwomtari-Nai	2	0	0	2	2
Northern Daly	2	1	1	2	2
Yawa-Saweru	2	1	1	2	2
Nyimang	1	1	2	2	2
Chimakuan	2	2	1	2	2
Palaihnihan	2	2	1	2	2
Chapacuran	2	1	2	2	2
Cahuapanan	2	1	2	2	2
Pauwasi	1	0	0	1	2
Peba-Yagua	1	1	2	2	2
Yuki-Wappo	1	2	2	2	2
Zamucoan	1	0	2	2	2
Kresh-Aja	2	1	2	2	2
Araucanian	1	1	1	1	1
Yokutsan	1	1	0	1	1
Turama-Kikori	1	0	0	1	1
Hatam-Mansim	1	0	0	1	1
West Bird's Head	1	0	0	1	1
Hibito-Cholon	0	0	1	1	1
Yeniseian	1	1	1	1	1
Amto-Musan	1	1	0	1	1
Ainu	1	1	1	1	1
Jicaquean	1	1	1	1	1
Gumuz	1	1	1	1	1
Huavean	1	1	1	1	1
Wintuan	1	1	1	1	1
Manubaran	1	1	0	1	1
Somahai	1	0	0	1	1
Kalapuyan	1	0	0	1	1
Piawi	1	1	0	1	1
Eastern Jebel	1	1	1	1	1
Mairasic	1	0	0	1	1
East Kutubu	1	0	0	1	1
Mombum-Koneraw	1	0	0	1	1
Doso-Turumsa	0	0	0	0	1
Dizoid	1	1	1	1	1
Limilngan-Wulna	1	0	0	1	1
Coosan	1	1	0	1	1
Nivkh	1	1	1	1	1
Chiquitano	1	0	1	1	1
Harakmbut	1	0	1	1	1
Garrwan	0	0	0	0	1
Ticuna-Yuri	1	1	1	1	1
Puri-Coroado	0	0	1	1	1
Koreanic	1	1	1	1	1
Rashad	1	0	0	1	1
Shastan	1	1	1	1	1
Bogia	1	0	0	1	1
Kawesqar	1	1	1	1	1
Kaure-Kosare	1	0	0	1	1
Katukinan	1	0	0	1	1
Tarascan	1	1	1	1	1
Taulil-Butam	1	0	0	1	1
Temeinic	1	1	1	1	1
Haida	1	1	1	1	1

Figure 5: Number of languages with non-empty `union` feature vectors in all language families.