URIEL+: Enhancing Linguistic Inclusion and Usability in a Typological and Multilingual Knowledge Base

Aditya Khan[†*], **Mason Shipton**[‡*], **David Anugraha**[†], **Kaiyao Duan**[†],
Phuong H. Hoang[†], **Eric Khiu**[§], **A. Seza Doğruöz**[&], **En-Shiun Annie Lee**[†‡]
[†]University of Toronto, Canada [‡]Ontario Tech University, Canada
[§]University of Michigan, USA [&]LT3, IDLab, Universiteit Gent, Belgium
{adityakhan,anugraha}@cs.toronto.edu, masonshipton25@gmail.com
as.dogruoz@ugent.be, annie.lee@ontariotechu.ca

Abstract

URIEL is a knowledge base offering geographical, phylogenetic, and typological vector representations for 7970 languages. It includes distance measures between these vectors for 4005 languages, which are accessible via the `lang2vec` tool. Despite being frequently cited, URIEL is limited in terms of linguistic inclusion and overall usability. To tackle these challenges, we introduce URIEL+,[1] an enhanced version of URIEL and `lang2vec` that addresses these limitations. In addition to expanding typological feature coverage for 2898 languages, URIEL+ improves the user experience with robust, customizable distance calculations to better suit the needs of users. These upgrades also offer competitive performance on downstream tasks and provide distances that better align with linguistic distance studies.

1 Introduction

The URIEL knowledge base and its `lang2vec` query tool (Littell et al., 2017) provide a standardized approach to representing languages as geographical, phylogenetic, and typological vectors. Geographical vectors contain distances between the locations where languages are spoken and 299 latitude/longitude coordinates. Phylogenetic and typological vectors consist of binary indicators that denote membership in language families or structural features, respectively. URIEL enables language comparisons through *language distance* calculations, which are performed using mathematical operations on these vectors.

Typological distance, or differences in language structure, is foundational for cross-linguistic comparisons (Haspelmath, 2023) and plays a crucial role in multilingualism, second language acquisition, and natural language processing (NLP) (Christina Nelson and Wrembel, 2021; Haspelmath, 2020). The challenge in defining *typological distance* lies in the structural uniqueness of each language, which complicates direct comparisons (Haspelmath, 2020). To navigate this complexity, linguists measure *typological distance* by focusing on particular linguistic domains (e.g., syntax, phonology, or phonemic inventory) (Nerbonne and Hinrichs, 2006).

Syntactic Distance Syntactic distance measures similarities and differences in grammatical structures using frameworks such as dependency trees and part-of-speech distributions. These methods provide quantitative comparisons of syntactic patterns between languages (Hammarström and O'Connor, 2013).

Phonological Distance Phonological distance measures similarities and differences in the overall sound systems of languages, including both segmental and suprasegmental features. This involves analyzing phonetic properties like voicing and place of articulation, as well as prosodic elements such as stress and intonation. Tools like n-gram models and phoneme frequency analysis position languages in a multidimensional space based on these comprehensive phonological characteristics (Gamallo et al., 2017).

Phonemic Inventory Distance Phonemic inventory distance measures the similarities and differences between the sets of phonemes in two languages, including both vowels and consonants. This involves comparing the number and types of phonemes present in each language, as well as their specific combinations, thereby providing insights into their phonemic structure to quantify how similar or different the phonemic inventories of the languages are (Bradlow et al., 2010).

While incorporating all of the domains mentioned above as feature categories within typologi-

[*] The authors contributed equally.
[1]The code is available at https://github.com/Masonshipton25/URIELPlus

Proceedings of the 31st International Conference on Computational Linguistics, pages 6937–6952
January 19–24, 2025. ©2025 Association for Computational Linguistics

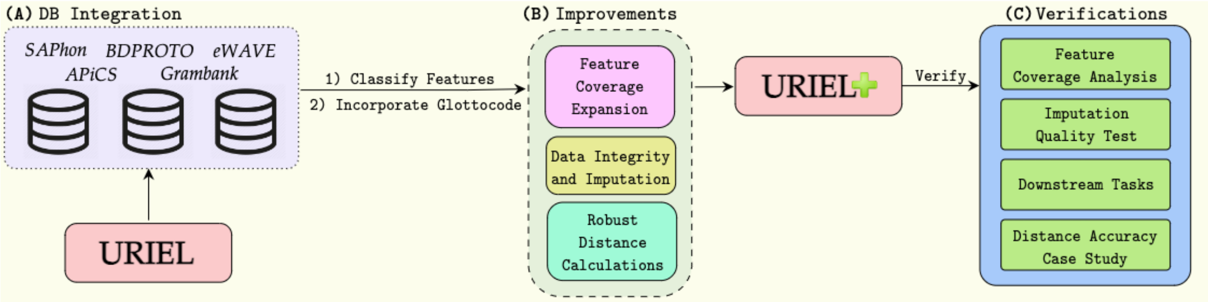

Figure 1: Overview of URIEL+. **(A)** We integrated five databases to incorporate more linguistic features across multiple languages, and **(B)** made several improvements on top of current URIEL. **(C)** We later verified the quality of URIEL+ through various experiments.

Language Pair	geo	gen	fea	syn	pho	inv
English-French	0.00	0.90	0.50	0.46	0.43	0.48
	0.02	**0.94**	**0.48**	**0.45**	**0.39**	**0.48**
Croatian-Serbian	0.0	0.13	0.90	0.71	0.83	0.68
	0.02	**0.00**	**0.00**	**0.00**	**NA**	**NA**

Table 1: Geographical (geo), genetic (gen), featural (typological) (fea), syntactic (syn), phonological (pho), and inventory (inv) distances for English-French in URIEL and URIEL+. **Bolded** entries indicate distances with URIEL+. **NA** entries indicate distances that could not be computed with URIEL+ (see Section 2.3).

cal vectors (featural vectors according to URIEL), URIEL aggregates linguistic data from multiple sources to provide a standardized, all-in-one measure of language distance. An example of URIEL distances can be seen in Table 1, where English and French are geographically identical, moderately distant typologically (syntactic, phonological, inventory), and genetically distant. This unified framework simplifies cross-linguistic comparisons by representing complex linguistic features as a single vector. By making these distances easily accessible, URIEL enables seamless integration into machine learning models, facilitating large-scale analysis and supporting diverse tasks, as outlined in Table 2.

Despite URIEL's established significance in measuring language distance, several areas for improvement have been identified regarding its feature coverage and usability (Toossi et al., 2024). For this reason, we introduce URIEL+, which aims to improve URIEL, with a focus on *typological features* and *typological distance*. Figure 1 outlines our contributions in enhancing both feature coverage and the overall usability of URIEL.

Feature Coverage Expansion Currently, 31% of the languages for which URIEL supports distance calculations have no data for any typological features, resulting in the use of undocumented default values (Toossi et al., 2024). These default values prevent meaningful differentiation between languages with missing data, rendering the calculated distances unreliable (Khiu et al., 2024). We addressed this issue by integrating additional databases into URIEL+ (Section 2.1). Incorporating these databases significantly enhances URIEL+'s feature coverage, increasing the number of languages available for featural distance calculations from 2724 to 4366.[2] Unlike the previous version, URIEL+ now includes morphological features, which are critical for representing morphologically rich languages (Samardzic et al., 2024).

Data Integrity and Imputation URIEL's use of default values for missing data without user awareness (Toossi et al., 2024) results in distances that may not be meaningful, especially when the calculation involves low resource languages. Researchers have mitigated this by developing their own imputation methods (Jin and Xiong, 2022) or by using URIEL's k-nearest-neighbor-imputed feature vectors[3] (Üstün et al., 2020; Glavaš and Vulić, 2021; Choenni et al., 2023), although the details and quality of URIEL's imputation remain undocumented. To address this issue, we integrated three well-evaluated imputation algorithms (Section 2.2) and provided users with the ability to choose their preferred imputation method. This functionality allows distance calculations for all languages in

[2] In URIEL, 4005 languages are available for distance calculation. However, only 2724 of these languages have actual data (the remaining 1281 rely on default values due to missing data).

[3] URIEL calls these vectors k-nearest-neighbour aggregated vectors.

NLP Task	Related Papers	# of citations
Cross-lingual transfer	Lin et al., 2019; Lauscher et al., 2020; Ruder et al., 2021	678
Dependency parsing	Üstün et al., 2020	108
Machine translation	Zhang and Toral, 2019; Li et al., 2024b; Dankers et al., 2022	160
Speech recognition	Adams et al., 2019; Zanon Boito et al., 2020	140
Performance prediction	Xia et al., 2020; Srinivasan et al., 2021; Khiu et al., 2024; Anugraha et al., 2024	65

Table 2: Natural language processing (NLP) tasks, related papers utilizing URIEL feature vectors and/or distances, and their total forward citation counts.

the knowledge base and enables users to select the most suitable method for their research needs.

Robust Distance Calculations with Confidence Scores Three issues affect `lang2vec` distance computations. First, distances are pre-computed, preventing modifications to feature vectors for updated distances. Second, Toossi et al., 2024 identified reproducibility issues, citing conflicting documentation on aggregation methods[4] and distance metrics. Third, `lang2vec` allows distance computations for languages without values known for both languages (shared data), leading to potentially meaningless results. Furthermore, calculations use all features across all URIEL data sources, limiting users to one-size-fits-all calculations. To meet specific needs, researchers have manually calculated distances between subsets (Papadimitriou and Jurafsky, 2020), concatenations (Adams et al., 2019), or both (Zhang and Toral, 2019; Hossain et al., 2020) of URIEL vectors. URIEL+ addresses these issues with a rigorous dynamic calculation system (Section 2.3), that allows customization of aggregation methods, metrics, features, and sources. URIEL's language distances often differ from linguistic measures. For example, URIEL shows Croatian and Serbian as distant languages (Table 1), although they should be similar (Samardzic et al., 2024). This mismatch could be indicative of poor data quality. Therefore, we introduce new confidence scores to assess the reliability of the distances between languages (Section 2.4).

To assess the improvement brought by the URIEL+ enhancements, we replicated three notable downstream usages of the original URIEL knowledge base (see Section 4.1): 1) LANGRANK for selecting transfer languages in cross-lingual learning (Lin et al., 2019); 2) LINGUALCHEMY for typological feature-driven language analysis (Adilazuarda et al., 2024); and 3) PROXYLM for performance prediction in multilingual settings (Anugraha et al., 2024). Our experimental results demonstrate that URIEL+ not only augments feature coverage and usability but also leads to better performance in practical NLP applications.

To assess how well URIEL+ distances align with linguistic distance metrics, we conducted a case study comparing typological distances between Central and South American Indigenous languages in URIEL and URIEL+ to a metric from Hammarström and O'Connor, 2013, which accounts for predictable traits, quirks, and historical dependencies (see Section 5.1). URIEL+ shows a higher correlation with the linguistic distances than URIEL, suggesting that URIEL+ is more aligned with linguistic measurements of typological distance.

2 From URIEL to URIEL+

This section outlines the expansions to URIEL and `lang2vec`. URIEL+ incorporates new databases to enhance feature coverage and implements imputation algorithms to handle missing values. Furthermore, `lang2vec` improves distance computations between languages and introduces new confidence scores to evaluate the reliability of these distances.

2.1 Integrating New Databases

URIEL+ includes five additional databases (Table 3), incorporating data for 2898 languages (2858 of which are low resource). Users can select databases (e.g., BDPROTO for ancient and reconstructed languages) based on their needs. By default, all five databases are included, providing data for 8071 languages, with 4366 capable of *typological distance* calculations. Data from the five databases was first preprocessed separately, then integrated altogether into URIEL to create URIEL+. We detail the process below.

[4] In `lang2vec`, distance is computed on data that aggregates feature information for each language (using union or average aggregation). Union aggregation sets each feature value to the max value across all sources, while average aggregation sets it to the average value.

Database	Reference	Contribution
SAPhon	Michael et al., 2015	Phonology and inventories for South American indigenous languages.
BDPROTO	Marsico et al., 2018	Phonology and inventories for ancient and reconstructed languages.
Grambank	Skirgård et al., 2023	Syntax and morphological data for ∼2500 languages.
APiCS	Mufwene, 2013	Typological data for pidgin and creole languages.
eWAVE	Kortmann et al., 2020	Syntax and morphological data for English dialects.

Table 3: Summary of database contributions and updates. **Bolded** entries highlight databases that have been newly added to URIEL+ and were not present in the original URIEL.

Binarization for Non-Binary Features While URIEL (and thus, URIEL+) supports only binary features, Grambank, APiCS, and eWAVE contain non-binary features, specifically, *nominal* and *ordinal variables*. *Nominal variables* are categorical variables with no inherent order, while *ordinal variables* represent different levels of a feature's presence in a language. To collect data from these databases, we first binarized the features. Nominal features were binarized using one-hot encoding, and ordinal features were binarized by creating a replacement feature that indicates whether that feature is present in the language.

Combining Redundant Features The Grambank, APiCS, and eWAVE databases contain features that overlap with those in URIEL or with each other, which can cause redundancy.

Suppose the values for a feature (say f_1) from one database can be inferred from another feature (f_2) in a different database. In such cases, we update the values in f_1 using inferred values from f_2, but not necessarily the other way around. For example, if a language has the URIEL feature "S_ARTICLE_WORD_BEFORE_NOUN," it necessarily has the Grambank feature "Are there prenominal articles?" Alternatively, if f_1 is equivalent to f_2, we infer values from f_2 and remove f_2 from its database. This reduces redundancy when we integrate the preprocessed datasets into URIEL.

Classifying and Renaming Features New features are classified as either syntactic, phonological, inventory, or morphological. This is implemented in the code by adding a prefix to the feature name. The prefixes are S_, P_, INV_, and M_ for syntactic, phonological, inventory, and morphological features, respectively - matching the conventions in the original URIEL knowledge base. In addition to adding these prefixes, we rename the feature names to align with the naming conventions in the original URIEL knowledge base. This involves capitalizing features, truncating their names, and replacing spaces with underscores.

Incorporating Glottocode Identifiers URIEL uses ISO 639-3 codes to identify languages. While these codes remain compatible with the updated SAPhon database,[5] the other databases now require glottocode identifiers (Forkel et al., 2022). Therefore, URIEL+ employs glottocodes to better support low resource languages, including those not covered by ISO 639-3, such as Eskimo Pidgin and Singlish. Outdated ISO 639-3 codes that coexisted with their replacements in URIEL, such as *gre* for Greek (now *ell*) and *alb* for Albanian (now *sqi*), have been removed to ensure up-to-date and unique language identifiers.

Summary of Implementation Details URIEL is structured as a three-dimensional matrix, with languages, features, and data sources as the three dimensions. To incorporate data from the new databases, we extend URIEL by adding new entries for corresponding languages, features, and sources. Initially, these new entries are marked as missing and are subsequently updated with the available data, ensuring that URIEL+ integrates new information while preserving existing data. For the updated SAPhon database, the new data replaces the missing values in the existing SAPhon data rather than creating any new feature columns.

2.2 Automatic Imputation Algorithms

Despite expanding URIEL and increasing its feature coverage, after combining the five databases into a single source, 87% of values remain missing. To address this, URIEL+ includes methods for

[5] The SAPhon database identifies languages with ISO 639-3 codes.

imputing missing feature data, enabling comprehensive distance calculations between languages even with incomplete data.

We provide several imputation algorithms in URIEL+, including k-NN imputation, which was also an option in URIEL. Additionally, we include MIDASpy (Lall and Robinson, 2023), a multiple imputation method implemented with deionizing autoencoders, and SoftImpute (Mazumder et al., 2010), which fits a low-rank matrix approximation via nuclear-norm regularization.

2.3 Robust Distance Calculations

We replaced pre-computed queries with a new function that *dynamically* computes distances based on current data in the knowledge base, allowing it to reflect database updates.

Since Toossi et al., 2024 found that union aggregation with angular distance produces distances most aligned with URIEL's, these are set as the default options for computing distances. To resolve the documentation ambiguity, we offer users the option to choose between union or average aggregation and between cosine or angular distances, enhancing flexibility and clarity. In addition, to address the issue of meaningless distances, we exclude languages without shared data from computations. Instead, we provide imputation algorithms for users who need these distances.

We offer a distance function that computes distance based on provided features rather than all available features under a provided feature category, allowing for any combination of features. Furthermore, users can specify a particular source to use data from rather than using aggregated data. For example, one could calculate the distance of languages using 49 syntactic features exclusively from the WALS source (Dryer and Haspelmath, 2013), as was done manually in Papadimitriou and Jurafsky, 2020.

These changes preserve all existing functionalities while introducing feature and source customizations in calculations.

2.4 Confidence Scores of Distance Calculations

To evaluate the quality of the calculated distances, confidence scores are often used, as suggested by prior studies (Salati et al., 2016; Bayram et al., 2023, 2024). These scores typically aggregate several key metrics, including: 1) the amount of inaccurate data (accuracy), 2) the proportion of missing data (completeness), 3) the agreement across different sources or adherence to established constraints (consistency), 4) the recency of the data (timeliness), and 5) the deviation from a reference distribution (skewness) (Bayram et al., 2023; Batini et al., 2009).

However, not all of these metrics are applicable or necessary in our case. Accuracy cannot be evaluated due to the absence of ground truth, timeliness is irrelevant since the data is non-temporal, and there is no natural reference distribution to measure skewness. Therefore, following Salati et al., 2016, our confidence scores focus solely on data completeness, data consistency, and a new metric we introduce: imputation quality.

Formally, given languages L_1 and L_2, we define data completeness $\mathcal{M}(L_1, L_2)$ as:

$$\mathcal{M}(L_1, L_2) = 1 - \frac{p(L_1) + p(L_2)}{2}$$

where $p(L_i)$ is the proportion of missing values for language L_i.

Next, recall there may be multiple sources of feature value j for language L_i. Let $\mathcal{S}_{i,j}$ denote the set of sources that provide values for feature j on language L_i and let $n_{i,j}$ be the cardinality of $\mathcal{S}_{i,j}$. Let $v_{i,j,s}$ denote the value of feature j for language L_i from source $s \in \mathcal{S}_{i,j}$. Define $m_{i,j}$ as the mode of the set $\{v_{i,j,s} : s \in \mathcal{S}_{i,j}\}$. We can then define $z_{i,j}$, the number of sources that agree with the mode $m_{i,j}$, as:

$$z_{i,j} = \sum_{s \in \mathcal{S}_{i,j}} \mathbb{1}\{v_{i,j,s} = m_{i,j}\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

Following this definition, we now define data consistency $\mathcal{C}(L_1, L_2)$ as:

$$\mathcal{C}(L_1, L_2) = \frac{a(L_1) + a(L_2)}{2}$$

where $a(L_i)$ is computed as:

$$a(L_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} \frac{z_{i,j}}{n_{i,j}}$$

with k_i representing the number of non-empty language features in URIEL+ for language L_i. If $k_i = 0$, we set $a(L_i) = 1$.

Finally, we define imputation quality $\mathcal{I}$ as a constant γ, where $\gamma \in [0, 1]$ depends on the metric chosen to define imputation quality. For specific numerical values of γ, see Table 4.

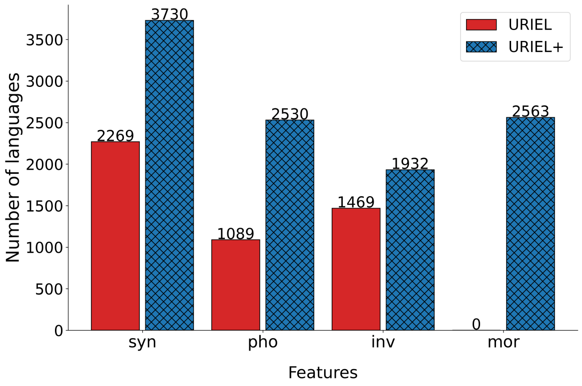

Figure 2: Number of languages with available syntactic (syn), phonological (pho), inventory (inv), and morphological (mor) data in URIEL and URIEL+ with all five databases.

3 Validating the Knowledge Base

Two of the ways we validated URIEL+ were through an analysis of feature coverage and imputation quality tests on the three algorithms.

3.1 Feature Coverage Analysis

Experimental Setup To compare feature coverage between URIEL and URIEL+, we calculated the number of languages in each that have available syntactic, phonological, inventory, and morphological data. For URIEL+, all five databases were integrated in this comparison. Additionally, we analyzed feature coverage by categorizing languages into high resource languages (HRLs), medium resource languages (MRLs), and low resource languages (LRLs), as defined by Joshi et al., 2020.

Results By integrating all five URIEL+ databases, the number of languages available for typological distance calculations increases from 4005 to 4366, representing a 9.01% increase. This expansion includes ancient, reconstructed, pidgin, creole, and dialectal languages. Figure 2 provides a detailed view of this expansion, highlighting that the number of languages with syntactic data increases from 2269 to 3730, a 64.39% increase, driven by the inclusion of Grambank, APiCS, and eWAVE. Phonological data coverage rises from 1089 to 2530, representing a 132.3% increase, facilitated by SAPhon, BDPROTO, APiCS. Inventory data experiences the smallest increase, expanding from 1469 to 1932, a 31.5% increase, with contributions from SAPhon, BDPROTO, and APiCS, which mainly augment existing data.

Figure 3 shows the breakdown of feature coverage by language resource level. For HRLs, URIEL+ now includes syntactic data for Arabic. The feature coverage for MRLs and LRLs improve substantially across all feature categories, with the most significant gains in phonology (70.8% more languages for MRLs and 134.5% more languages for LRLs). These increases are expected, given that SAPhon, BDPROTO, and APiCS focus on LRLs and contribute substantial new phonological data. In addition, LRLs see a large increase in syntax, with a 65.6% increase in the number of languages with syntactic information compared to URIEL. With these feature coverage expansions, URIEL+ provides more comprehensive distance calculations to more languages.

Method	Union-Agg				Average-Agg	
	Accuracy	Precision	Recall	F1	RMSE	MAE
Mean	0.8024	0.7248	0.5656	0.6354	0.3597	0.2608
MIDASpy	0.8435	0.7819	0.6737	0.7238	0.3302	0.2171
k-NN	0.8678	0.8136	**0.7338**	0.7717	0.3069	**0.1809**
SoftImpute	**0.8875**	**0.8801**	0.7300	**0.7980**	**0.2883**	0.1886

Table 4: Summary of imputation quality test results, with metrics grouped by union-aggregated and average-aggregated data. We maximise F1 on union-aggregated data and minimise RMSE on average-aggregated data. For k-NN, we choose $k = 9$ for union-aggregated data and $k = 15$ for average-aggregated data. **Bolded** entries indicate the best results in each category.

3.2 Imputation Quality Test

Experimental Setup To evaluate the three imputation algorithms and validate our choice of imputation algorithm for downstream tasks, we used the imputation quality test from Li et al., 2024a. This test involves removing 20% of non-missing data, imputing it, and comparing the predictions to known values using metrics like F1 for binary data and root mean square error (RMSE) for continuous data. Since the URIEL paper (Littell et al., 2017) did not include detailed metrics and procedures for its k-NN aggregation, the imputation quality test focuses solely on URIEL+ and does not compare it with URIEL. Imputation was performed on aggregated data (union or average), with missing dialect data filled using the parent language's data, which typically has similar typological features. More details on the data used and the methodology of the quality test can be found in Appendix A.1.

Results For both imputed union and average data, all imputation algorithms outperform the mean imputation as our baseline across all metrics (Table 4).

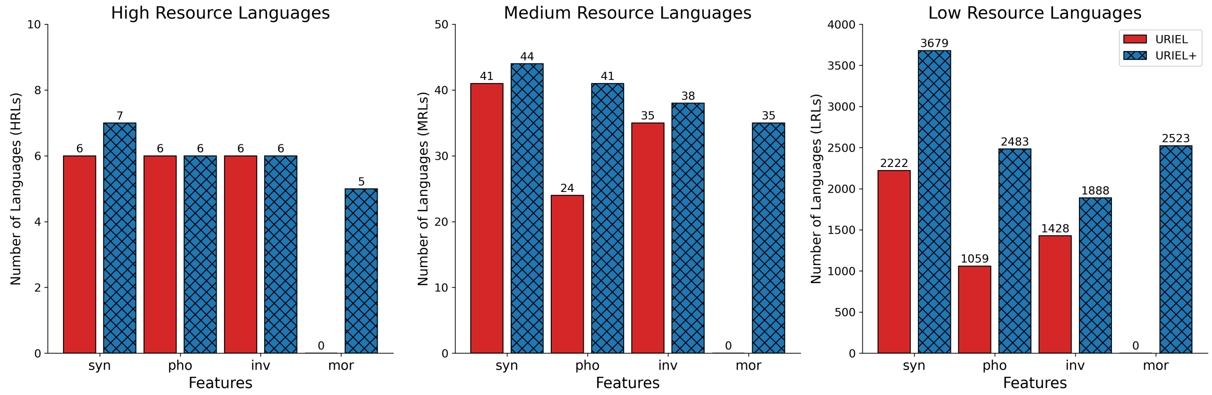

Figure 3: Number of languages with available syntactic (syn), phonological (pho), inventory (inv), and morphological (mor) data in URIEL and URIEL+ with all five databases, is shown for high resource languages (HRLs), medium resource languages (MRLs), and low resource languages (LRLs) (Joshi et al., 2020) from left to right.

Comparing the performance on union-aggregated data, all algorithms have at least 6% higher precision compared to the baseline. However, for each algorithm, recall is always worse than precision, meaning that each algorithm is good at imputing binary values of 1 but is too conservative in doing so. The algorithm that performs the best at maximizing both precision and recall is SoftImpute, as it has the highest F1 score.

On average-aggregated data, all algorithms have lower RMSEs and MAEs than the baseline. However, SoftImpute is the most preferable of these algorithms. This is because SoftImpute performs almost the same as k-NN at minimizing errors overall since its MAE exceeds k-NN's by only 0.0057. However, it does much better at minimizing *large* errors (it has the lowest RMSE).

Since SoftImpute performed the best, we used URIEL+ with union source aggregation and this method in both downstream tasks experiments (except for LINGUALCHEMY where we used union- and average-aggregation) and the linguistic case study. A more granular analysis of how these imputation methods fare on specific feature categories can be found in Appendix A.2.

4 URIEL+ on Downstream Tasks

As shown in Table 2, URIEL has been used in various downstream NLP tasks. Building on this, we apply URIEL+ to different NLP tasks by comparing its performance with URIEL distances and vectors across three frameworks (LANGRANK, LINGUALCHEMY, PROXYLM). These frameworks are employed to evaluate multiple NLP tasks.

4.1 Experimental Setup

LANGRANK (Lin et al., 2019) LANGRANK predicts cross-lingual transfer languages using multiple data-related features, including features from all six URIEL distance categories. LANGRANK is evaluated on part-of-speech tagging (POS), machine translation (MT), dependency parsing (DEP), and entity linking (EL) using top-3 Normalized Discounted Cumulative Gain (NDCG@3), and shows higher average scores than other baselines.

LINGUALCHEMY (Adilazuarda et al., 2024) LINGUALCHEMY employs a regularization technique that utilizes URIEL's syntactic and geographic vectors to guide language representations in pre-trained models. The evaluation was conducted on three tasks: semantic relatedness using SemRel2024 (Ousidhoum et al., 2024), news classification using MasakhaNews (Adelani et al., 2023), and intent classification using MASSIVE (FitzGerald et al., 2023). Semantic relatedness was assessed using Pearson correlation, while intent and news classification were measured by accuracy. LINGUALCHEMY improves performance on mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), benefiting LRLs and unseen language generalization.

PROXYLM (Anugraha et al., 2024) PROXYLM estimates the performance of language models in multilingual NLP tasks, using proxy models and URIEL distances without fine-tuning, saving time and computational resources. PROXYLM outperforms state-of-the-art performance prediction methods in terms of RMSE, while also demonstrating robustness and efficiency. PROXYLM was eval-

Downstream Tasks	Subtasks	URIEL	URIEL+
LANGRANK	POS	22.9	24.6 (↑ 7.42%)
	MT	42.55	**45.45** (↑ 6.82%)
	DEP	71.7	72.85 (↑ 1.60%)
	EL	66.50	66.80 (↑ 0.45%)
LINGUALCHEMY	SemRel2024	0.008	**0.012** (↑ 50.00%)
	MasakhaNews	76.13	**76.43** (↑ 0.39%)
	MASSIVE	72.95	**73.16** (↑ 0.29%)
PROXYLM	Unseen	4.35	**4.08** (↑ 6.21%)
	Random	2.92	2.82 (↑ 3.42%)
	LOLO	3.84	3.85 (↓ 0.26%)

Table 5: Results for each subtasks in LANGRANK, LINGUALCHEMY, and PROXYLM. **URIEL** refers to tasks using the URIEL knowledge base, and **URIEL+** to tasks using the URIEL+ knowledge base. Higher values indicate better performance for LANGRANK and LINGUALCHEMY, while lower values do so for PROXYLM. ↑ indicates an improvement, and ↓ signifies a decline relative to URIEL. **Bolded** numbers indicate results that are statistically significant at a 0.05 significance level using the Wilcoxon signed-rank test (Rey and Neuhäuser, 2014). More detail about the metrics used along with the breakdown of the results can be found in Appendices B.1 and B.2, respectively.

uated on M2M100 (Fan et al., 2020) and NLLB (Costa-jussà et al., 2022) under three settings: Unseen (generalization on unseen languages to the pre-trained language models), Random (random split), and LOLO (Leave-One-Language-Out).

For each of the frameworks, we will refer to each of the tasks, datasets, or settings evaluations as "subtasks" for convenience. Further details on the experimental settings for each of the frameworks along with the downstream tasks can be found in Appendix B.1.

4.2 Results

Table 5 compares the summary of URIEL and URIEL+ on the three downstream tasks. LANGRANK with URIEL+ distances consistently outperforms URIEL distances across all subtasks. The most significant performance gain is observed in the POS subtask, followed by MT, DEP, and EL (detailed breakdown in Table 11 in the Appendix), which compares LANGRANK using only language features (referred to as LANGRANK (lang feats)) and both dataset and language features (referred to as LANGRANK (all)).

Both LANGRANK (lang feats) and LANGRANK (all) show performance gains in the POS and MT subtasks, with a notable boost in MT from including dataset features. This is because MT often emphasizes data-specific characteristics over linguistic features, resulting in a larger improvement for LANGRANK with all data and language features rather than with only language features (Lin et al., 2019). For the EL subtask, LANGRANK (all) performs slightly worse, likely due to the absence of dataset-specific information. Interestingly, using only language features improves performance for the EL subtask. In the DEP subtask, integrating both dataset and language features enhances performance, reinforcing the value of combining URIEL+ features with dataset-specific information (Lin et al., 2019).

In LINGUALCHEMY, URIEL+ shows a slight performance increase across all subtasks, highlighting its advantage. Table 12 in the Appendix indicates that average syntax vectors yield mixed results, with both increases and decreases in performance compared to syntax k-NN vectors. This suggests that syntax k-NN vectors, which used union-aggregation, retain more feature diversity by including any available presence. Meanwhile, average aggregation might lose key distinctions since it smooths out differences between features. Therefore, k-NN vectors are more reliable when incorporating URIEL+ features for LINGUALCHEMY.

For PROXYLM, competitive performance is observed, with significant improvements in Unseen and Random settings, as detailed in Tables 5 and 13 in the Appendix. These results suggest that URIEL+ features improve generalization to previously unseen languages. The insignificant change in performance for the LOLO setting may indicate that the existing URIEL features across multiple languages are already sufficient for maintaining robust predictions when leaving one language out from the regressor.

The Wilcoxon signed-rank test (Rey and Neuhäuser, 2014) indicates that most results in Table 5 are statistically significant (p-values less than 0.05), except for PROXYLM's Random and LOLO settings, and LANGRANK's POS, EL, and DEP tasks. The non-significance of PROXYLM's LOLO setting suggests that the small performance drop is negligible. In contrast, the non-significance of LANGRANK's POS, EL, and DEP tasks is likely due to the limited number of target languages in LANGRANK.

Overall, URIEL+ demonstrates competitive performance across all tasks compared to URIEL, especially where language-specific features are crucial.

Language Pair	URIEL	URIEL+	G_d
Paez-Bintucua	0.50	0.55	0.49
Sambu-Cayapa	0.50	0.54	0.48
Ulua-Paez	0.70	0.74	0.45
Sumo-Paez	0.70	0.74	0.45
Paez-Misquito	0.80	0.58	0.43
Quiche-Paez	0.60	0.64	0.43
Quiche-Boruca	0.80	0.65	0.41
Quiche-Huaunana	0.60	0.61	0.40
Xinca-Cofan	0.70	0.73	0.39
Xinca-Boruca	0.90	0.79	0.39
Quiche-Colorado	0.60	0.64	0.38
Quiche-Cayapa	0.60	0.62	0.36
Quiche-Cuna	0.80	0.56	0.35
Paya-Bintucua	0.60	0.50	0.35
Cuna-Boruca	0.00	0.51	0.35
Paya-Muisca	0.60	0.51	0.33
Huaunana-Boruca	0.80	0.51	0.32
Paya-Cagaba	0.50	0.43	0.31
Xinca-Camsa	0.70	0.76	0.30
Quiche-Lenca	0.70	0.73	0.28
Rank Correlation with G_d	-0.05	**0.19**	N/A

Table 6: Language distances from URIEL, URIEL+ and the dependency-sensitive Gower coefficient (G_d), and rank correlation of URIEL and URIEL+ with G_d.

5 Distance Alignment Case Study

In this section, we demonstrate that the distances provided by URIEL+ align more closely with a linguistic distance metric. This is illustrated through a case study on distance measures between two languages, as conducted by Hammarström and O'Connor, 2013.

5.1 Experimental Setup

We assessed the accuracy of URIEL and URIEL+ distance measures by comparing them to a modified Gower coefficient, G_d (Gower, 1971), for typological distance (Hammarström and O'Connor, 2013). This metric weights scores based on both predictable and idiosyncratic traits, while accounting for dependencies and historical contact. The case study focused on Isthmo-Colombian languages, which now have more data in URIEL+, due to updates in the SAPhon database. We evaluated which knowledge base aligns better with G_d by using Kendall's rank correlation. This approach was chosen over direct comparisons with G_d distances due to differences in unit scales, opting instead for unit-agnostic rank correlation. To compare distances from URIEL and URIEL+ with those derived from the coefficient, we computed featural distances using `lang2vec`.[6]

5.2 Results

Table 6 shows the distances for each language pair in the case study, where Kendall's rank correlations between G_d and the distances from URIEL and URIEL+ are -0.05 and 0.19 respectively.

We used the Perm-Both hypothesis test (Deutsch et al., 2021) to compare the significance between these two correlations. The p-value was found to be 0.307 which is not statistically significant at the 0.05 level. The lack of significance is likely due to the small sample size, as smaller datasets tend to have higher p-values (Johnson, 1999). Comparing additional language pairs could help further validate these findings. Although the difference in correlation is not statistically significant, URIEL+ shows a trend toward better alignment with G_d, suggesting that it may improve the alignment of typological distance metrics in NLP with those used in linguistics.

6 Conclusion

We introduce URIEL+, an enhanced knowledge base that expands the coverage of typological features by integrating five additional databases, providing data for 2858 LRLs. Furthermore, URIEL+ improves the robustness and usability of `lang2vec` distances through carefully selected imputation methods, a rigorous study of appropriate distance calculations, and the establishment of new confidence scores to validate distance reliability. In addition, we demonstrate URIEL+'s competitive performance on downstream NLP tasks and its closer alignment with real-world linguistic distances through a case study. These improvements are critical for a wide range of multilingual applications and contribute to the linguistic inclusion of LRLs. As an open-source tool, we hope the community will contribute to its ongoing improvements and database expansions in the future.

Limitations

With the new features from Grambank, APiCS, and eWAVE databases, URIEL+ predominantly emphasizes syntactic (474 features) and morphological (133 features) data, with fewer contributions to phonological (30 features) and inventory (163 features) data. This shift skews the focus of featural

[6]The ISO 639-3 codes and glottocodes required for these distances are provided in the Appendix, Table 10

distance towards grammar (607 features overall) over sound, potentially underrepresenting phonological aspects (193 features overall). To address this imbalance, we plan to include more phonological features, creating more specific distinctions based on current features in future work.

Similar to URIEL, URIEL+ does not have information on language scripts. To address this, we will introduce scripts as a feature category using ScriptSource (Holloway, 2013), which covers 8290 languages in future work.

Other future work would be integrating URIEL+ as an external knowledge base with large language models, merging structured knowledge with flexible language modeling.

Acknowledgement

We thank Hasti Toossi and Guo Qing Huai for their valuable feedback. We also thank Dr. Patrick Littell (NRCC/Canada) for his insightful discussion and suggestions.

References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2401.06034*.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2024. Proxylm: Predicting language model performance on multilingual tasks via proxy models. *arXiv preprint arXiv:2406.09334*.

Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3).

Firas Bayram, Bestoun S. Ahmed, and Erik Hallin. 2024. Adaptive data quality scoring operations framework using drift-aware mechanism for industrial applications. *Journal of Systems and Software*, 217:112184.

Firas Bayram, Bestoun S. Ahmed, Erik Hallin, and Anton Engman. 2023. DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications. In *EASE '23: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering*, Oulu, Finland. Association for Computing Machinery.

Ann Bradlow, Cynthia Clopper, Rajka Smiljanic, and Mary Ann Walter. 2010. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, 52(11):930–942. Non-native Speech Perception in Adverse Conditions.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Halina Lewandowska Christina Nelson, Iga Krzysik and Magdalena Wrembel. 2021. Multilingual learners' perceptions of cross-linguistic distances: a proposal for a visual psychotypological measure. *Language Awareness*, 30(2):176–194.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Robert Forkel et al. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6):917–924.

Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

J. C. Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Harald Hammarström and Loretta O'Connor. 2013. *Dependency-sensitive typological distance*, pages 329–352. Walter de Gruyter.

Martin Haspelmath. 2020. The structural uniqueness of languages and the value of comparison for language description. *Asian Languages and Linguistics*, 1:346–366.

Martin Haspelmath. 2023. *Word-Class Universals and Language-Particular Analysis*, pages 15–40. Oxford University Press.

Steph Holloway. 2013. Scriptsource - writing systems, computers and people.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics.

Renren Jin and Deyi Xiong. 2022. Informative language representation learning for massively multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5158–5174, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Douglas H. Johnson. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63(3):763–772.

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, Online. Association for Computational Linguistics.

Eric Khiu, Hasti Toossi, Jinyu Liu, Jiaxu Li, David Anugraha, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*.

Ranjit Lall and Thomas Robinson. 2023. Efficient multiple imputation for diverse data in python and r: Midaspy and rmidas. *Journal of Statistical Software*, 107:1–38.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

JiaHang Li, ShuXia Guo, RuLin Ma, Jia He, XiangHui Zhang, DongSheng Rui, YuSong Ding, Yu Li, LeYao Jian, Jing Cheng, and Heng Guo. 2024a. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, 24(1):41.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

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.

Egidio Marsico, Sebastien Flavier, Annemarie Verkerk, and Steven Moran. 2018. BDPROTO: A database of phonological inventories from ancient and reconstructed languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.

Michael, Lev, Tammy Stark, Emily Clem, and Will Chang. 2015. South american phonological inventory database v2.1.0.

Salikoko S. Mufwene. 2013. Atlas of pidgin and creole language structures online.

John Nerbonne and Erhard Hinrichs. 2006. Linguistic distances. In *Proceedings of the Workshop on Linguistic Distances*, pages 1–6, Sydney, Australia. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 13 languages. *Preprint*, arXiv:2402.08638.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Denise Rey and Markus Neuhäuser. 2014. Wilcoxon-signed-rank test. In M. M. Atkinson and W. J. Piegorsch, editors, *International Encyclopedia of Statistical Science*, pages 1658–1659. Springer. First Online: 01 January 2014.

Donald B. Rubin. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michele Salati, Pierre-Emmanuel Falcoz, Herbert Decaluwe, Gaetano Rocco, Dirk Van Raemdonck, Gonzalo Varela, Alessandro Brunelli, and on behalf of the ESTS Database Committee. 2016. The european thoracic data quality project: An aggregate data quality score to measure the quality of international multi-institutional databases. *European Journal of Cardio-Thoracic Surgery*, 49(5):1470–1475.

Tanja Samardzic, Ximena Gutierrez, Christian Bentz, Steven Moran, and Olga Pelloni. 2024. A measure for transparent comparison of linguistic diversity in multilingual NLP data sets. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3367–3382, Mexico City, Mexico. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Hasti Toossi, Guo Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz, and En-Shiun Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the URIEL knowledge base. In *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, Mexico City, Mexico. Association for Computational Linguistics.

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.

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, Online. Association for Computational Linguistics.

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France. European Language Resources Association.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Imputation Quality Test

In this section, we describe more details about the experimental setup and results of the imputation quality test.

A.1 Experimental Settings

Imputation on our data is always performed on aggregated data, either union or average. Before imputation, we allow users to fill in missing dialect data with parent language data, assuming typological similarity (e.g., filling American English with English data). This method is applied across all imputation quality tests and downstream tasks (see section B.1).

We follow the imputation quality test from Li et al. (2024a), where 20% of non-missing data is randomly removed, imputed, and the imputed values y_{pred} are compared to the original values y_{true} using a metric $\mathcal{L}$. Metrics $\mathcal{L}$ include accuracy, precision, recall, and F1 for union data (binary), and RMSE/MAE for average data (continuous). We optimize RMSE for average data to penalize outliers and F1 for union data due to class imbalance (more 0s). The method that optimizes $\mathcal{L}$ is selected as the best imputation approach.

We compare k-NN, MIDASpy, and SoftImpute against mean imputation as a baseline. Note that two additional steps are used in this methodology, depending on the imputation method.

k-NN Imputation We perform a hyperparameter search over $k = 3, 6, 9, 12, 15$ using 5-fold cross-validation to select k, optimizing $\mathcal{L}$, prior to comparing k-NN with other imputation methods.

MIDASpy Imputation MIDASpy performs multiple imputations, pooling results based on Rubin's rules (Rubin, 1987). We report the average $\mathcal{L}$ across 5 imputed datasets. The network is initialized with 3 layers and 256 units per layer, following Lall and Robinson (2023).

A.2 Detailed Results

A more granular analysis of imputation performance by feature type is shown in Tables 7 and 8. SoftImpute consistently performs best across

Imputation Method	Feature Type	Accuracy	Precision	Recall	F1
Mean	mor	0.7428	0.6456	0.2914	0.4016
	syn	0.7386	0.6491	0.5445	0.5922
	inv	0.8975	0.8453	0.6970	0.7640
	pho	0.8426	0.8174	0.8397	0.8284
MIDASpy	mor	0.7962	0.6937	0.5582	0.6186
	syn	0.8034	0.7580	0.6397	0.6939
	inv	0.9101	0.8484	0.7577	0.8005
	pho	0.8574	0.8492	0.8328	0.8409
k(9)-NN	mor	0.8388	0.7611	**0.6644**	**0.7095**
	syn	0.8263	0.7830	**0.6942**	0.7359
	inv	0.9301	0.8806	0.8173	0.8477
	pho	0.8566	0.8547	0.8231	0.8386
SoftImpute	mor	**0.8485**	**0.8259**	0.6188	0.7075
	syn	**0.8475**	**0.8518**	0.6808	**0.7568**
	inv	**0.9444**	**0.9249**	**0.8342**	**0.8772**
	pho	**0.9376**	**0.9657**	**0.8938**	**0.9284**

Table 7: Summary of feature type (syntactic (syn), phonological (pho), inventory (inv), and morphological (mor)) analysis of imputation quality test over union-aggregated data. **Bolded** entries indicate the best results in each category.

Imputation Method	Feature Type	RMSE	MAE
Mean	mor	0.4103	0.3370
	syn	0.4089	0.3352
	inv	0.2636	0.1428
	pho	0.3464	0.2442
MIDASpy	mor	0.3771	0.2814
	syn	0.3688	0.2695
	inv	0.2516	0.1278
	pho	0.3227	0.2120
k(15)-NN	mor	0.3471	**0.2315**
	syn	0.3482	**0.2339**
	inv	0.2234	**0.0965**
	pho	0.3219	0.1857
SoftImpute	mor	**0.3398**	0.2485
	syn	**0.3344**	0.2451
	inv	**0.1987**	0.1024
	pho	**0.2296**	**0.1447**

Table 8: Summary of feature type (syntactic (syn), phonological (pho), inventory (inv), and morphological (mor)) analysis of imputation quality test over average-aggregated data. **Bolded** entries indicate the best results in each category.

all feature types. Notably, all methods perform far better on inventory and phonological data than on syntactic or phonological data. In particular, for SoftImpute, there is at least a $+7\%$ gap between 1) inventory and phonological data, and 2) syntactic and morphological data. This performance gap correlates with the proportion of missing values in each feature type. Specifically, we see that:

- 85.73% of morphological data is missing
- 91.25% of syntactic data is missing
- 77.06% of inventory data is missing
- 81.88% of phonological data is missing

Notably, syntactic and morphological data have the highest proportions of missing values. This suggests that the imputation algorithms perform better when when there are fewer missing values to impute.

Remark The imputation quality γ in Section 2.4 can be the F1 score for union-aggregated data or $1 - \text{RMSE}$ for average-aggregated data.

B Downstream Tasks

In this section, we describe more details about the experimental setup and results for downstream tasks.

Language	URIEL code	URIEL+ code
Albanian	alb	alba1267
Arabic	ara	stan1318
Azerbaijani	aze	nort2697
Chinese	zho	mand1415
Estonian	ekk	esto1258
Malay	msa	stan1306
Oromo	orm	east2652
Persian	fas	west2369
Swahili	swa	swah1253

Table 9: Languages used in downstream task experiments with ISO 639-3 codes in URIEL but without equivalent glottocodes in URIEL+, and the glottocodes that were used as replacements.

B.1 Experimental Settings

To run LANGRANK and PROXYLM experiments with URIEL+ distances, we simply replaced the URIEL distances with the new URIEL+ distances.

While most languages used for downstream tasks had ISO 639-3 codes with corresponding glottocodes, some did not, often because the ISO 639-3 codes were outdated. In these cases, we assigned the glottocode of the most appropriate language from URIEL+. The languages without glottocodes and their replacements are listed in Table 9.

No GPU was required for LANGRANK and PROXYLM experiments. All experiments for LIN-

LINGUALCHEMY were run on two Tesla V100 32GB GPUs.

Experimental Settings for LANGRANK We could not completely replicate the original LANGRANK baselines (Lin et al., 2019) due to unclear parameter specifications. We attempted to keep most of the experimental setup the same as the original paper, except for the following:

- When assigning relevance to languages that have the same BLEU or accuracy score, the lowest rank in the group is used, as it made the most sense and produced good results.
- For calculation of Normalized Discounted Cumulative Gain (NDCG), we attempted to use the original paper's formula where the Discounted Cumulative Gain is defined as

$$\mathrm{DCG@}p = \sum_{i=1}^{p} \frac{2_i^{\gamma} - 1}{\log_2(i+1)}.$$

However, an alternative formulation of DCG, defined as

$$\mathrm{DCG@}p = \sum_{i=1}^{p} \frac{\gamma_i}{\log_2(i+1)},$$

was chosen instead, as it produced baseline results much closer to the original paper.

Experimental Settings for LINGUALCHEMY We could not find the source code for the vectors for the MasakhaNews (Adelani et al., 2023) news classification dataset and the SemRel2024 (Ousidhoum et al., 2024) semantic relatedness dataset, as well as a pipeline for SemRel2024. To address this, we created vectors with both URIEL and URIEL+ feature data for these datasets and constructed a pipeline for SemRel2024, using the approach applied to the MASSIVE (FitzGerald et al., 2023) intent classification dataset and MasakhaNews.

Only new syntactic vectors needed to be created, as the geography vectors for all languages in the LINGUALCHEMY datasets remained unchanged. Despite their names, the syntax_knn and syntax_average vectors used SoftImpute with union and average aggregation, respectively, rather than k-NN, since SoftImpute was employed for imputation in all other downstream task experiments.

A $10\times$ URIEL loss scaling factor was used as it provided the best results in Adilazuarda et al., 2024.

Language	URIEL code	URIEL+ code
Sambu	emp	nort2972
Cayapa	cbi	chac1249
Paya	pay	pech1241
Bintucua	arh	arhu1242
Cágaba	kog	cogu1240
Ulua	sum	sumu1234
Paez	pbb	paez1247
Sumo	sum	sumu1234
Cuna	cuk	sanb1242
Boruca	brn	boru1252
Muisca	chb	chib1270
Huaunana	noa	woun1238
Misquito	miq	misk1235
Quiche	quc	kich1262
Lenca	len	lenc1239
Xinca	xin	xinc1237
Camsa	kbh	cams1241
Cofan	con	cofa1242
Colorado	cof	colo1256

Table 10: Languages used in the distance accuracy case study with corresponding identifiers. The table includes **URIEL code** using **ISO 639-3** identifiers and **URIEL+ code** using **glottocode** identifiers.

Experimental Settings for PROXYLM The proxy regressor in PROXYLM had two different datasets: an English-centric dataset borrowed from the MT560 dataset (Gowda et al., 2021) and a Many-to-many languages dataset borrowed from the NUSA dataset (Cahyawijaya et al., 2023). The Many-to-many languages dataset contained the language Batak which does not have an ISO 639-3 code. Anugraha et al., 2024 used the code "bhp" for Bima, a Batak family member, which has a glottocode equivalent: "bima1247". We used the Ensemble model, which performed best for PROXYLM. The hyperparameters of the proxy regressor were the same as in Anugraha et al., 2024.

B.2 Detailed Results

Tables 11, 12, and 13 show the results for using URIEL+ on LANGRANK, LINGUALCHEMY, and PROXYLM respectively.

Method	Machine Translation		Entity Linking		POS Tagging		Dependency Parsing	
	URIEL	**URIEL+**	**URIEL**	**URIEL+**	**URIEL**	**URIEL+**	**URIEL**	**URIEL+**
LangRank (lang feats)	35.2	35.4 (↑ 0.20)	64.7	66.4 (↑ 1.70)	18.3	20.4 (↑ 2.10)	73.2	70.4 (↓ 2.8)
LangRank (all)	49.9	55.5 (↑ 5.60)	68.3	67.2 (↓ 1.10)	27.5	28.8 (↑ 1.30)	70.2	75.3 (↑ 5.10)
Avg. LangRank	42.55	45.45 (↑ 2.90)	66.50	66.80 (↑ 0.30)	22.9	24.6 (↑ 1.70)	71.7	72.85 (↑ 1.15)

Table 11: The results for two LangRank model (Lin et al., 2019) on predicting cross-lingual transfer measured using average NDCG@3 multiplied by 100 (**higher is better**). LangRank (lang feats) considers only language vectors, while LangRank (all) denotes LangRank with language vectors and additional dataset-dependent features such as size and type-token ratio.

Model	Feature Type	MASSIVE		MasakhaNews		SemRel2024	
		URIEL	**URIEL+**	**URIEL**	**URIEL+**	**URIEL**	**URIEL+**
mBERT	syntax_avg	66.01	65.71 (↓ 0.30)	70.96	71.52 (↑ 0.56)	0.005	0.009 (↑ 0.004)
	syntax_avg+geo	65.59	65.74 (↑ 0.15)	70.77	70.75 (↓ 0.02)	0.01	0.008 (↓ 0.002)
	syntax_knn	65.83	65.85 (↑ 0.02)	71.08	72.07 (↑ 0.99)	0.014	0.012 (↓ 0.002)
	syntax_knn+geo	65.45	66.00 (↑ 0.55)	70.91	71.36 (↑ 0.45)	0.008	0.013 (↑ 0.005)
	syntax_knn+syntax_avg	65.45	66.25 (↑ 0.80)	71.22	70.52 (↓ 0.70)	0.013	0.015 (↑ 0.002)
	syntax_knn+syntax_avg+geo	65.59	65.82 (↑ 0.23)	70.99	71.13 (↑ 0.14)	0.01	0.012 (↑ 0.002)
XLM-R	syntax_avg	80.51	80.47 (↓ 0.04)	81.31	80.91 (↓ 0.40)	0.007	0.012 (↑ 0.005)
	syntax_avg+geo	80.20	80.35 (↑ 0.15)	81.58	81.71 (↑ 0.13)	0.015	0.013 (↓ 0.002)
	syntax_knn	80.11	80.53 (↑ 0.42)	81.18	81.40 (↑ 0.22)	0.017	0.013 (↓ 0.004)
	syntax_knn+geo	80.00	80.33 (↑ 0.33)	81.32	81.90 (↑ 0.58)	0.005	0.007 (↑ 0.002)
	syntax_knn+syntax_avg	80.47	80.19 (↓ 0.28)	80.63	82.00 (↑ 1.37)	0.012	0.014 (↑ 0.002)
	syntax_knn+syntax_avg+geo	80.17	80.39 (↑ 0.22)	81.58	81.89 (↑ 0.31)	0.008	0.010 (↑ 0.002)
Avg.		72.95	73.16 (↑ 0.21)	76.13	76.43 (↑ 0.30)	0.008	0.012 (↑ 0.004)

Table 12: The results for LinguAlchemy (Adilazuarda et al., 2024) using URIEL loss scale of 10 obtained by averaging the accuracy for MASSIVE and MasakhaNews datasets and the Pearson correlation for SemRel2024 dataset across all languages under different benchmarks (**higher is better**). **URIEL** denotes LinguAlchemy with URIEL vectors, while **URIEL+** denotes LinguAlchemy with URIEL+ vectors.

Dataset	**Experimental Setting**	**M2M100**		**NLLB**	
		URIEL	**URIEL+**	**URIEL**	**URIEL+**
English Centric	Random	3.64 $\pm$ 0.19	3.62 $\pm$ 0.18 (↓ 0.02)	3.80 $\pm$ 0.37	3.79 $\pm$ 0.39 (↓ 0.01)
	LOLO	3.90 $\pm$ 0.22	3.84 $\pm$ 0.22 (↓ 0.06)	4.14 $\pm$ 0.23	4.10 $\pm$ 0.23 (↓ 0.04)
	Unseen	4.35 $\pm$ 0.25	4.08 $\pm$ 0.25 (↓ 0.27)	NA	NA
Many-to-Many	Random	2.47 $\pm$ 0.35	2.36 $\pm$ 0.29 (↓ 0.11)	1.76 $\pm$ 0.42	1.49 $\pm$ 0.32 (↓ 0.27)
	LOLO	3.64 $\pm$ 0.24	3.67 $\pm$ 0.24 (↑ 0.03)	3.67 $\pm$ 0.18	3.79 $\pm$ 0.26 (↑ 0.12)
Avg.		3.60	3.51 (↓ 0.09)	3.34	3.29 (↓ 0.05)

Table 13: The results for ProxyLM (Anugraha et al., 2024) using XGBoost Ensemble in average RMSE $\pm$ standard deviation under different datasets and settings (**lower is better**). **URIEL** denotes ProxyLM with URIEL as its language features, while **URIEL+** denotes ProxyLM with URIEL+ as its language features.