Simple Additions, Substantial Gains: Expanding Scripts, Languages, and Lineage Coverage in URIEL+

Mason Shipton[♦] **York Hay Ng**[♥] **Aditya Khan**[♥] **Phuong Hanh Hoang**[♥]
Xiang Lu[♠] **A. Seza Doğruöz**[♣] **En-Shiun Annie Lee**[♦♥]
[♦]Ontario Tech University [♥]University of Toronto
[♠]University of Michigan [♣]LT3, IDLab, Universiteit Gent
masonshipton25@gmail.com, as.dogruoz@ugent.be,
annie.lee@ontariotechu.ca

Abstract

The URIEL+ linguistic knowledge base supports multilingual research by encoding languages through geographic, genetic, and typological vectors. However, data sparsity (e.g. missing feature types, incomplete language entries, and limited genealogical coverage) remains prevalent. This limits the usefulness of URIEL+ in cross-lingual transfer, particularly for supporting low-resource languages. To address this sparsity, we extend URIEL+ by introducing script vectors to represent writing system properties for $7,488$ languages, integrating Glottolog to add $18,710$ additional languages, and expanding lineage imputation for $26,449$ languages by propagating typological and script features across genealogies. These improvements reduce feature sparsity by 14% for script vectors, increase language coverage by up to $19,015$ languages $(1,007\%)$, and boost imputation quality metrics by up to 35%. Our benchmark on cross-lingual transfer tasks (oriented around low-resource languages) shows occasionally divergent performance compared to URIEL+, with performance gains up to 6% in certain setups.

Keywords: URIEL knowledge base, multilingual NLP, low-resource languages, linguistic typology

1. Introduction

Linguistic knowledge bases enable us to understand the relationships between languages, in particular conducing the rise of cross-lingual transfer learning (Bjerva et al., 2020). One such knowledge base is URIEL (Littell et al., 2017), which represents languages with three vector types: geographic (i.e. based on 299 coordinate points), genetic (i.e. family membership), and typological (i.e. syntactic, phonological, and inventory features). Through the lang2vec tool, URIEL provides pre-computed language distances. Downstream applications of these distances include cross-lingual transfer (Lin et al., 2019), dependency parsing (Üstün et al., 2020), and performance prediction (Khiu et al., 2024, Anugraha et al., 2025).

URIEL+ (Khan et al., 2025) partially addresses data sparsity and transparency issues in URIEL (Littell et al., 2017) as identified by Toossi et al. (2024), by expanding data coverage to over $8,000$ languages. Furthermore, URIEL+ prevents distance computation between language pairs that lack shared features. Although this is restrictive, URIEL+ allows users to circumvent this by advanced imputation methods to fill in missing values (e.g. by way of the SoftImpute algorithm, due to Mazumder et al. (2010)). *Lineage imputation*, a method of partial imputation, optionally populates features for dialects of six languages (Spanish, French, English, German, Malay, and Arabic), given that phylogenetic relationships of languages are strong predictors of language similarity (Dunn et al., 2011).

Nonetheless, three key limitations persist in URIEL+: (1) the absence of writing system representations, (2) incomplete language coverage (i.e. many languages catalogued in Glottolog (Hammarström and Forkel, 2022), a widely used catalogue of the world's languages and language families, remain absent), and (3) the narrow scope of lineage imputation. Each limitation reflects a form of data sparsity (i.e. missing feature types, missing languages, and limited phylogenetic propagation). Our study addresses all limitations by extending URIEL+ through (1) *script vectors*, (2) *Glottolog integration*, and (3) *expanded lineage imputation* (see Figure 1 which summarises our three additions that collectively reduce sparsity and expand feature coverage in URIEL+).

Table 1 shows language pairs that are expected to exhibit similarity in script but divergence in linguistic typology. Each pair shares a common script (e.g. Cyrillic for Russian and Kazakh, Latin for English, Turkish, Spanish, and Finnish). Korean and Japanese do not currently share the same script. However, both languages historically incorporated Chinese characters, which should result in low script distances. Before and after imputation, URIEL+ produces the anticipated pattern of low script distances coupled with higher typological distances, confirming that shared script does not necessarily imply typological similarity and that script as a feature provides unique insights into languages that typological currently does not.

Contribution 1: Script Vectors While URIEL+ captures a wide range of linguistic dimensions, it omits information about writing systems. Script properties are necessary for encoding structural and functional infor-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11045–11059
11–16 May 2026.

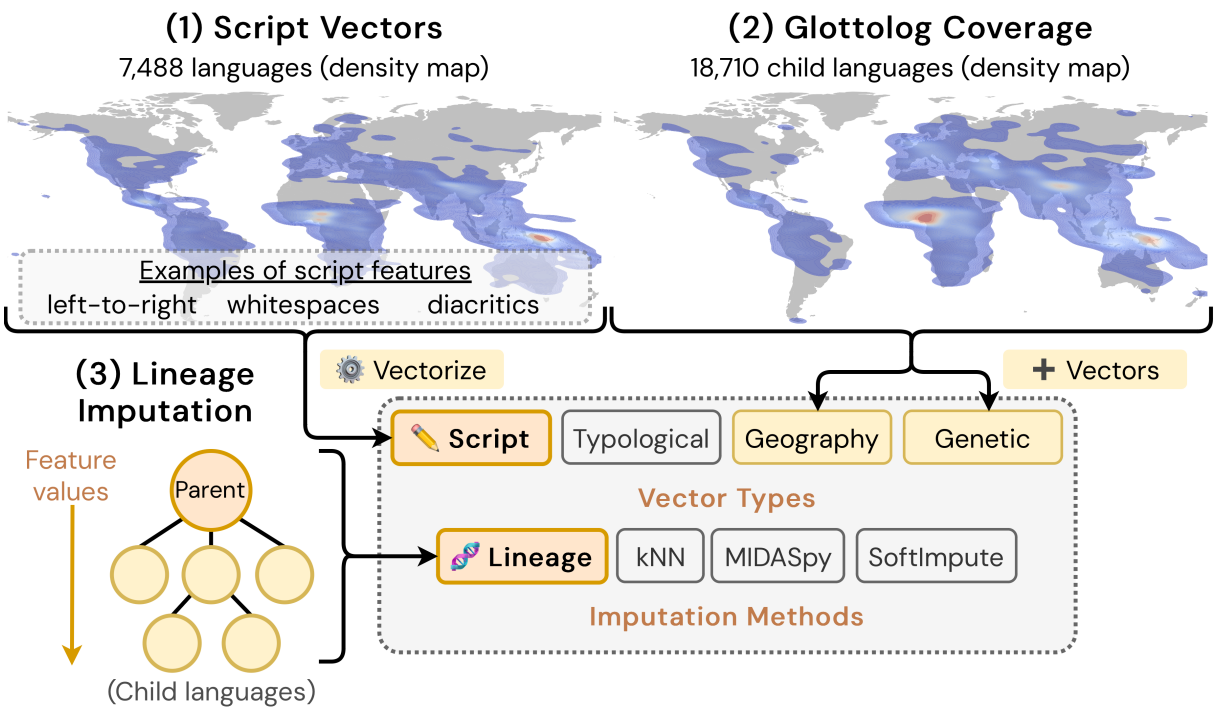

Figure 1: Overview of our additions to mitigating data sparsity in URIEL+. (1) Density map of $7,488$ languages with integrated script vectors. (2) Density map of the $18,710$ Glottolog child languages added to URIEL+. (3) Illustration of expanded lineage imputation, where missing values in child languages are filled from ancestors.

Language Pair	Glottocodes	Pre-Imputation		Post-Imputation	
		Script	Typology	Script	Typology
Russian - Kazakh	russ1263 - kaza1248	0.4788	0.6310	0.5000	0.6650
Korean - Japanese	kore1280 - nucl1643	0.2380	0.4706	0.2380	0.4827
English - Turkish	stan1293 - nucl1301	0.1864	0.6001	0.1864	0.6031
Spanish - Finnish	stan1288 - finn1318	0.0000	0.4949	0.0000	0.5054

Table 1: Languages similar in script (i.e. script distance closer to 0) but divergent in typology (i.e. featural distance closer to 1), shown before and after union aggregation of databases, and SoftImpute imputation.

mation relevant to natural language processing (NLP) tasks. Differences in script types (e.g. alphabetic or abjadic[1]) result in challenges in tokenisation and morphological processing. Moreover, shared scripts often reflect historical or cultural contact between genetically distant languages (e.g. Arabic script for Persian and Urdu). An encoding of these relationships is missing in URIEL+, but it is important for cross-lingual transfer (Zhuang et al., 2025). Therefore, we introduce *script vectors* to encode structural and social properties of writing systems for $7,488$ languages. Through analyses in feature coverage and sparsity, we highlight how vectors expand URIEL+'s coverage in a novel direction, particularly benefitting low-resource languages where script vectors are among the few reliably documented vector types.

[1]A writing system that represents individual sounds as symbols.

Contribution 2: Glottolog Integration A second form of sparsity arises from limited language coverage in URIEL+. While URIEL+ includes over $8,000$ languages, Glottolog (Hammarström and Forkel, 2022) catalogues $18,710$ additional languages with well-documented genealogical information. By incorporating these missing languages, URIEL+ contains a complete encoding of language families in Glottolog. Beyond increasing coverage, this integration enables a phylogeny-based method for imputing missing values, discussed below.

Contribution 3: Expanded Lineage Imputation Lineage imputation in URIEL+ fills missing values in child languages by copying feature values from their immediate parent. As it stands, it supports only 6 medium- to high-resource parent languages. This reflects the empirical observation that related languages tend to share typological features (Dunn

et al., 2005; Jäger and Wahle, 2021). We extend this approach to systematically propagate typological and script features from parent languages to all child languages across the knowledge base, forming an *expanded lineage imputation* method that benefits $26,449$ languages in total following Glottolog integration. Furthermore, we demonstrate that this method yields consistent performance gains in imputation quality.

Finally, we benchmark the impact of these contributions in choosing languages for cross-lingual transfer across a variety of tasks, a common use case of linguistic knowledge bases. The performance gains from ablations of our contributions highlight their practicality for cross-lingual transfer, while particularly supporting low-resource languages.

2. Literature Review

Current research shows strong support for our three contributions to URIEL+. The existing literature in NLP demonstrates the importance of script data for modelling language relationships, while linguistic studies show that genetically related languages tend to share typological structures. Together, these findings support our proposed additions, which aim to reduce data sparsity and improve coverage in URIEL+.

2.1. Script Data Usage in Literature

The script of a language is a first-order determinant of language similarity, and it is crucial in many NLP tasks.

For *morphological inflection*, Murikinati et al. (2020) show that transfer between related languages degrades when scripts differ. Transliterating the source into the target script (or both into a common script) improves accuracy, sometimes outweighing typological relatedness. For *transfer learning*, Amrhein and Sennrich (2020) and Tufa et al. (2024) report that romanising the script of a child language improves transfer performance when parent and child use different scripts. For *multilingual pre-training*, LANGSAMP (Liu et al., 2025) adds learnable script embeddings to the transformer output during masked language modelling (MLM), yielding better zero-shot transfer and more reliable source-language selection. Xhelili et al. (2024) further show that transliteration-based post-training alignment improves *cross-lingual alignment*. Without such alignment, token representations from different scripts are nearly linearly separable. Finally, Zhuang et al. (2025) find that script differences systematically hinder transfer in low-resource languages.

Prior approaches rely on ad hoc, language-specific pre-processing (e.g. romanisation per Amrhein and Sennrich (2020)) which are unevenly available and biased toward high-resourced languages. These issues with bias can be addressed by adding script vectors to URIEL+, which reduces reliance on pre-processing and improves modelling of linguistic similarity, especially for low-resource languages.

2.2. Glottolog as a Language Resource

Despite URIEL+'s extensive coverage of over $8,000$ languages, many languages remain missing. For example, Bajjika (Toossi et al., 2024) is missing in URIEL and URIEL+, while Western Armenian and Sakizaya are missing in URIEL [2] (Novotný et al., 2021). The missing languages highlight the need for a more comprehensive resource to ensure inclusive representation in linguistic knowledge bases.

Glottolog provides a solution to this sparsity by offering a comprehensive, authoritative catalogue of the world's languages, dialects, and language families (Hammarström and Forkel, 2022), encompassing approximately $18,000$ more languages than URIEL+. It uses a standardised system of language codes (Glottocodes) and assigns language codes to each language family and language, including dialects. URIEL+ uses these codes to integrate data from linguistic sources such as BDPROTO (ancient languages; (Marsico et al., 2018)), Grambank (Skirgård et al., 2023), and eWAVE (English dialects; (Kortmann et al., 2020)). Moreover, Glottolog provides evidence-based genealogical classifications, facilitating integration across linguistic datasets. Recognised as the authoritative source on the world's languages, Glottolog has been used in high-impact NLP and computational linguistics research (Liang et al., 2023; Bommasani et al., 2022), demonstrating its value for addressing gaps in URIEL+ and supporting more inclusive multilingual studies.

2.3. Phylogenetic Feature Propagation

Phylogenetic relationships of languages are strong predictors of structural similarity. Dunn et al. (2005) demonstrate that typological features carry a phylogenetic signal in Oceanic and Papuan languages. Dunn et al. (2011) consider evolutionary rates of typological and lexical features, showing that slower-changing features track lineage more closely. Jäger and Wahle (2021) proposed methods to estimate typological feature distributions while controlling for shared ancestry, and Hübler (2022) examined structural features, identifying which exhibit strong phylogenetic signal.

These existing studies motivate us to extend URIEL+'s lineage imputation to infer missing typological and script features for all languages based on their parent languages. This approach reduces sparsity and enhances coverage across URIEL+.

[2]While Sakizaya has been added to URIEL+ via the Grambank linguistic source (Skirgård et al., 2023), Western Armenian remains absent from URIEL+

ScriptSource Feature	Possible Values	Binarised URIEL+ Feature(s)
Type	Alphabet; Abjad; Abugida; Featural; Logo-syllabary; Syllabary	`SC_ALPHABET`, `SC_ABJAD`, `SC_ABUGIDA`, `SC_FEATURAL`, `SC_LOGO_SYLLABARY`, `SC_SYLLABARY`
Case	Yes; No	`SC_CASE`
Ligatures	Required; Optional; None	`SC_LIGATURES`, `SC_REQUIRED_LIGATURES`

Table 2: Three examples of how binarisation of ScriptSource features was performed using one-hot encoding.

3. Method

This section describes how we extend URIEL+ to reduce data sparsity and broaden feature coverage via three steps: (1) introducing script vectors, (2) integrating languages from Glottolog, and (3) expanding lineage imputation.

3.1. Adding Script Vectors

We introduce *script vectors* as a new vector type in URIEL+ to represent structural and social properties of writing systems. Features were sourced from ScriptSource (Holloway, 2013), a database of scripts, their typologies, and language usage. ScriptSource data is publicly viewable but it is not directly accessible. However, with assistance from the one of the ScriptSource administrators, we obtained dataframes describing: (a) writing system properties for all scripts, (b) languages in ScriptSource with their ISO 639-3 codes, and (c) scripts available for each language.

The writing system properties from dataframe (a) were binarised as shown in Table 2. For example, the ScriptSource feature *Type*, which includes values such as *Alphabet*, *Abjad*, *Abugida*[3]; was expanded such that each value became its own binary feature (e.g. `SC_ALPHABET`, `SC_ABJAD`, `SC_ABUGIDA`), with a value of 1 (if present) or 0 (if absent). This binarisation process is the same as it was done in URIEL (Littell et al., 2017) and URIEL+ (Khan et al., 2025) for typological vectors. For languages that can be written in multiple scripts (e.g. Kazakh uses Cyrillic, Arabic, and Latin scripts), the union of feature values was applied. This was implemented as a logical OR operation, where a feature was marked as present (1) if it existed in any of the language's associated scripts. As an example, Kazakh is marked as having both the `SC_ABJAD` feature (from Arabic) and the `SC_ALPHABET` feature (from Cyrillic and Latin).

These script vectors were integrated alongside URIEL+'s genetic, geographic, and typological vectors. Language-to-language distances based on script features were computed using normalised angular distance:

$$D_{\text{ang}}(A, B) = \frac{1}{\pi} \arccos\left(\frac{A \cdot B}{||A||||B||}\right)$$

where lower values indicate higher similarity (e.g. French and Spanish using the Latin script). This addition fills a representational gap, providing a high-coverage source of data for low-resource languages in particular.

3.2. Full Integration of Glottolog

We integrate all $18,710$ remaining Glottolog languages into URIEL+, expanding genealogical and geographic coverage. More specifically, we retrieved a list of all languages from Glottolog in their Glottocodes. Languages from Glottolog that were already in URIEL+ (over $8,000$ of them) were excluded. Glottolog was turned into a linguistic source in URIEL+ (Khan et al., 2025), similar to the World Atlas of Language Structures (WALS; (Dryer and Haspelmath, 2013)), Syntactic Structures of the World's Languages (SSWL; (Koopman, 2009)), and Grambank (Skirgård et al., 2023), where it can be integrated by choice to add the new languages. Unlike the other linguistic sources, Glottolog does not include new typological or script features, and it does not include typological data. Although these languages have no typological and script data, nearly all of them possess genetic classifications, allowing them to benefit directly from the lineage-based imputation to reduce sparsity.

Geographic coordinates and family affiliations for the Glottolog languages were obtained using the `lingtypology` package (Moroz, 2017), as URIEL+ did for their new languages. Although Glottolog contributes no new vector types, its inclusion provides the genealogical backbone required for large-scale feature propagation of URIEL+.

3.3. Lineage Imputation Methodology

To further reduce sparsity, we extend URIEL+'s *lineage imputation* to all languages, applying it to both typological and script vectors. This approach assumes that parent languages and their children share highly similar structural and orthographic properties, a pattern empirically supported in URIEL+ (89.8% agree-

[3]A writing system where consonants carry inherent vowels.

ment in feature values between parent and children for typological data, 97.0% for script data). These high agreement rates motivate a systematic propagation of known feature values to fill gaps across languages. However, this propagation does not resolve all missing values. Instead, it serves as a structured preprocessing step that can complement subsequent imputation methods such as SoftImpute (Mazumder et al., 2010), where otherwise the imputation of empty language vectors yields an uninformative result.

Algorithm 1 outlines the expanded lineage imputation algorithm. Parent-child mappings were derived from Glottolog (Hammarström and Forkel, 2022), containing $13,372$ parent languages which support imputation for $26,449$ child languages after Glottolog integration. Relationships were modelled as a forest of language trees. Imputation was performed via breadth-first search (BFS). Feature values from parent nodes were propagated to immediate children (continuing until all descendants were filled). This ensures hierarchical, consistent propagation, improving URIEL+ coverage and reducing sparsity at every phylogenetic level. The lineage imputation algorithm can be enabled or disabled if required, as described in the documentation.

Algorithm 1 Expanded Lineage Imputation via Breadth-First Search (BFS)

Construct a mapping of Glottocodes from parent to child languages.
Define root languages as all parent languages who have no parent.
for all root language r **do**
    Initialise queue $q \leftarrow [r]$ and visited set $s \leftarrow \{r\}$.
    while q not empty **do**
        Dequeue language p.
        for all child language c of p **do**
            if $c \notin s$ **then**
                Let f_p, f_c be binary feature vectors for p, c.
                For all i, set $f_c[i] \leftarrow f_p[i]$ if $f_c[i] =$ NaN and $f_p[i] \neq$ NaN.
                Enqueue c to q and add c to s.
            end if
        end for
    end while
end for

4. Analysis of Knowledge Base Statistics

This section examines how our additions affect the structure and completeness of URIEL+. We focus on (1) feature coverage (to measure the breadth of linguistic data across feature types and resource levels), (2) sparsity (to assess the proportion and distribution of missing values before and after expanded lineage imputation), and (3) correlations between script distances and other language distances (to evaluate the distinctiveness of script-based information). This analysis summarises how we make URIEL+ less sparse, thereby augmenting the availability of data for downstream applications.

4.1. Low-Resource Languages Have Higher Feature Coverage

We measured feature coverage in URIEL+ before and after expanded lineage imputation, counting the number of languages with at least one feature with data for syntactic, phonological, inventory, morphological, and script features. All five linguistic sources in URIEL+ and Glottolog languages were integrated. Languages were categorised as high-, medium-, or low-resource (HRL, MRL, LRL) following Joshi et al. (2020).

Figure 2 shows that script vectors substantially improve baseline coverage, providing data for $7,488$ languages before expanded lineage imputation (which is nearly double that of syntactic vectors ($3,730$ languages)). This difference is driven by gains in LRLs coverage. Syntactic vectors cover 7 HRLs, 44 MRLs, and $3,679$ LRLs. Meanwhile, script vectors cover 6 HRLs ($\downarrow 17\%$), 43 MRLs ($\downarrow 2\%$), and $7,439$ LRLs ($\uparrow 102\%$). Although script vectors include marginally fewer HRLs and MRLs, they reach nearly twice as many LRLs, highlighting their capacity to alleviate sparsity where documentation is most limited.

Expanded lineage imputation amplifies these gains for LRLs across all vector types. Morphological vectors show the smallest absolute increase, adding $6,805$ languages ($\uparrow 270\%$) due to sparse documentation. Syntactic vectors gain $9,512$ ($\uparrow 259\%$) languages. Phonological, inventory, and script vectors exhibit the largest gains, adding $19,030$ ($\uparrow 766\%$), $19,015$ ($\uparrow 1,007\%$), and $12,701$ ($\uparrow 171\%$) languages, respectively, consistent with their genealogical stability. All three vector types now cover over $20,000$ languages, demonstrating that inheritance-based propagation substantially enhances URIEL+'s representation of low-resource languages.

4.2. Lineage Imputation Lowers Sparsity

A significant limiting factor to the reliability of URIEL+ language vectors is their sparsity (i.e. the proportion of missing values). Breaking down typological vectors (into syntax, morphological, phonological and inventory vectors), Figure 3 demonstrates the sparsity in language vectors by type, before expanded lineage imputation. While our findings confirm that the high sparsity in typological vectors is consistent in URIEL+ across types ($78\% - 92\%$), script vectors notably possess a sparsity of only 14%. Coupled with their strong feature coverage, their low sparsity highlights the comprehensive nature of the data provided by script vectors. Furthermore, comparing

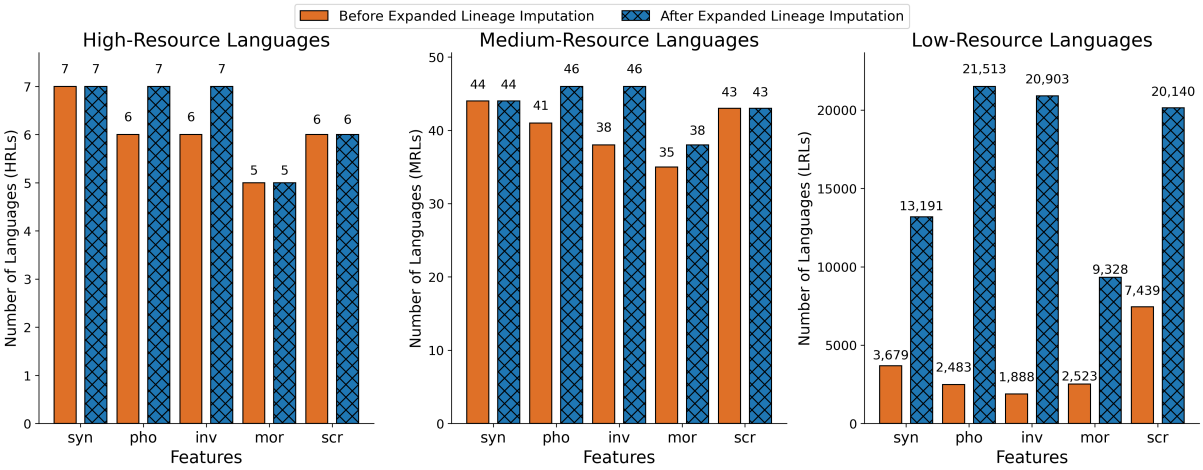

Figure 2: Number of languages with available syntactic (syn), phonological (pho), inventory (inv), morphological (mor), and script (scr) data in URIEL+ before and after expanded lineage imputation. All linguistic sources in URIEL+, as well as Glottolog languages, are integrated. Shown for high-resource (HRL), medium-resource (MRL), and low-resource (LRL) languages (Joshi et al., 2020) from left to right.

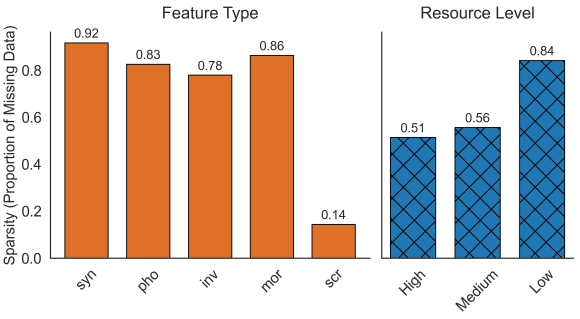

Figure 3: Sparsity of language vectors by type (syntax, phonological, inventory, morphological, script) and by resource level, **before** expanded lineage imputation.

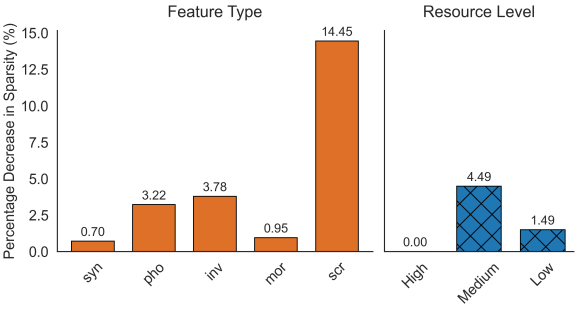

Figure 4: Percentage decrease in sparsity in language vectors by type (syntax, phonological, inventory, morphological, script) and by resource level, **after** expanded lineage imputation.

the overall sparsity in typological and script vectors across the resource levels of languages, we find that data sparsity is skewed, as sparsity in LRLs (84%) is substantially higher than in MRLs (56%) and HRLs (51%). The relatively high sparsity even among HRLs is mainly due to the large number of features included in URIEL+. URIEL+ has 800 features, many of whose values are derived from only one of the fourteen underlying data sources. Consequently, numerous features are sparsely populated across languages. For example, several features (e.g. S_REGULARIZED_REFLEXIVES_PARADIGM, S_DEMONSTRATIVES_FOR_DEFINITE_ARTICLES, and S_LIKE_AS_A_QUOTATIVE_PARTICLE) originate from the eWAVE (Kortmann et al., 2020) database and have recorded values for only around 70 languages out of the thousands represented in the knowledge base.

Our expanded lineage imputation strategy provides a reliable method for filling missing entries with linguistically grounded values in languages. Figure 4 shows the subsequent reduction in sparsity. This manifests in a reduction in sparsity (see Figure 4), particularly with sparsity in script vectors decreasing by 14%, i.e. to a sparsity of only 12%. However, comparing reductions in sparsity by vector type and resource level, expanded lineage imputation yields greater sparsity reductions where vectors are less sparse. In particular, sparsity reductions are minimal in syntax vectors and for LRLs, where sparsity was highest originally. This highlights how its efficacy is tied to the original sparsity of vectors since expanded lineage imputation relies on knowing feature values in parent languages.

4.3. Scripts Provide Unique Information

To justify the inclusion of script data, script distances should provide a different signal than typological, genetic, and geographic distances. In particular, they should not be substantially correlated with any of the other distance measures within URIEL+. Otherwise, script data would be redundant for the purpose of distance calculations–a primary use case of URIEL+.

To consider how similar script distances are to other distances in URIEL+, we study the correlation of a matrix containing all script distances with all other distance matrices in URIEL+. A distance matrix is a symmetric matrix, such that the ijth entry of the matrix is the distance between languages i and j in URIEL+. One can compute the Spearman correlation between the script matrix and any of the other distance matrices (Spearman, 1904). To test the significance of the coefficient, a two-sided Mantel test [4] is conducted.

Mantel tests assume the lack of any kind of structured dependence (e.g. phylogenetic relationships between languages) between the two matrices (Guillot and Rousset, 2013). This is clearly not the case for language distances. Languages within the same language family are expected to have smaller distances than those with languages outside their family (see Dunn et al. (2011)). To mitigate the violation of this assumption, we perform block permutation (999 many), where each permutation block (Winkler et al., 2015) is the top-level language family for a language, as reported in Glottolog. This alters the permutation procedure to allow us to account for phylogenetic dependencies between language distances. In total, there are 592 language family blocks (361 of which are singletons) and approximately 95% of languages fall into non-trivial language families (which mitigates discreteness caused by singletons). These tests are run on union-aggregated language vectors. For imputation, we run the correlation analysis with and without expanded lineage imputation, before using SoftImpute. We do not integrate any additional languages from Glottolog (i.e. beyond the $8,171$ already in URIEL+) because they do not have script data. The significance threshold we set is $\alpha = 0.05/7 \approx 0.007$. We apply a Bonferroni correction to account for conducting multiple hypothesis tests (Dunn, 1961).

Table 3 demonstrates that the magnitude of the correlation between script distances and all other distances in URIEL+ is at most 0.119, which is very minimal. This is irrespective of whether expanded lineage imputation was used or not. However, as seen in Table 3, statistical significance is more common without expanded lineage imputation. This suggests that lineage imputation essentially acts as a phylogenetic regulariser that gives SoftImpute fewer degrees of freedom to infer patterns that might induce weak correlation with script distances.

Distance	ρ (No Lineage)	ρ (Lineage)
gen	0.119	0.114
inv	0.059	0.057
mor	−0.026	−0.035
pho	**−0.041**	−0.046
syn	**0.060**	0.051
typ	**0.041**	0.023
geo	**−0.095**	**−0.095**

Table 3: Spearman correlation (ρ) between language and script distance measures. "Lineage" denotes using expanded lineage imputation prior to SoftImpute. Significant results are **bolded**.

Even if some correlations are significant, all correlations are minimal. Taking geographic (post expanded lineage imputation) as an example, $\rho \approx -0.095$ is marginal. This means that the coefficient of variation, $R^2 \approx 0.009$, meaning that geography distances explain less than 1% of the statistical variation in script distances. This supports our claim that script distances provide information that is largely *orthogonal* to existing distance information within URIEL+, and hence, provide unique information.

5. Evaluating our Additions

We validated our URIEL+ additions through two complementary evaluations: (a) testing imputation quality using the SoftImpute and lineage imputation algorithms and (b) applying language distances in cross-lingual transfer tasks using LangRank.

5.1. Imputation Quality Test

Experimental Setup To assess the contribution of script vectors and expanded lineage imputation to reducing sparsity, we conducted two imputation quality tests following the methodology used in URIEL+ (Khan et al., 2025) and Li et al. (2024). The procedure removes 20% of known (i.e. non-missing) data, imputes these values, and compares predictions to the ground truth using F1 for binary features and root mean square error (i.e. RMSE) for continuous features. For our first test, we evaluated the SoftImpute algorithm. SoftImpute previously achieved the strongest performance of the three URIEL+ imputation algorithms and was chosen for downstream and distance-alignment experiments (Khan et al., 2025). For this test, languages were pre-filled via our expanded lineage imputation procedure, however, filled values are not held out during imputation evaluation. This choice is to solely identify the quality of SoftImpute. With that said, we also report the results when

[4] That is, a permutation test that builds a distribution of correlation statistics based on computing the Spearman correlation coefficient on permuted distance matrices (Mantel, 1967).

Stage	Vector	Union-Agg				Average-Agg	
		Accuracy	Precision	Recall	F1	RMSE	MAE
− Lineage	typ	0.8821	0.8767	0.7092	0.7841	0.2909	0.1914
	syn	0.8410	0.8496	0.6650	0.7461	0.3357	0.2467
	pho	0.9140	0.9454	0.8472	0.8936	0.2570	0.1678
	inv	0.9432	0.9267	0.8245	0.8726	0.2015	0.1040
	mor	0.8451	0.8295	0.5998	0.6962	0.3374	0.2473
	scr	0.9939	0.9949	0.9826	0.9887	0.0769	0.0248
+ Lineage	typ	0.8880 (↑ 0.67%)	0.8821 (↑ 0.62%)	0.7238 (↑ 2.06%)	0.7951 (↑ 1.40%)	0.2845 (↓ 2.20%)	0.1876 (↓ 1.99%)
	syn	0.8511 (↑ 1.20%)	0.8660 (↑ 1.93%)	0.6781 (↑ 1.97%)	0.7606 (↑ 1.94%)	0.3262 (↓ 2.83%)	0.2399 (↓ 2.76%)
	pho	0.9048 (↓ 1.01%)	0.9180 (↓ 2.90%)	0.8540 (↑ 0.80%)	0.8848 (↓ 0.98%)	0.2626 (↑ 2.18%)	0.1691 (↑ 0.77%)
	inv	0.9436 (↑ 0.04%)	0.9242 (↓ 0.27%)	0.8294 (↑ 0.59%)	0.8742 (↑ 0.18%)	0.2034 (↑ 0.94%)	0.1074 (↑ 3.27%)
	mor	0.8545 (↑ 1.11%)	0.8319 (↑ 0.29%)	0.6312 (↑ 5.24%)	0.7178 (↑ 3.10%)	0.3260 (↓ 3.38%)	0.2366 (↓ 4.33%)
	scr	**0.9946** (↑ 0.07%)	**0.9957** (↑ 0.08%)	**0.9842** (↑ 0.16%)	**0.9899** (↑ 0.12%)	**0.0740** (↓ 3.77%)	**0.0242** (↓ 2.42%)

Table 4: Performance of SoftImpute on typological and script vectors before (−) and after (+) expanded lineage imputation. Metrics are reported for union-aggregated and average-aggregated evaluation schemes. Values that are imputed through lineage imputation are excluded from being held out during imputation evaluation. **Bolded** values indicate the best score for each metric.

Vector	Union-Agg				Average-Agg	
	Accuracy	Precision	Recall	F1	RMSE	MAE
typ	0.9886	0.9823	0.9757	0.9790	0.1068	0.0114
syn	0.9925	0.9902	0.9892	0.9897	0.0863	0.0075
pho	0.9879	0.9863	0.9846	0.9854	0.1099	0.0121
inv	0.9870	0.9768	0.9634	0.9701	0.1139	0.0130
mor	0.9937	0.9856	0.9948	0.9902	0.0793	0.0063
scr	**0.9987**	**0.9976**	**0.9974**	**0.9975**	**0.0366**	**0.0013**

Table 5: Performance of lineage imputation on typological and script vectors. Metrics are reported for union-aggregated and average-aggregated evaluation schemes. **Bolded** values indicate the best score for each metric.

we allow lineage-imputed values to be held out during imputation evaluation (i.e. treated as known), and the results of this can be found in Table 7 in Appendix A.

For our second test, we evaluated solely the lineage imputation, skipping SoftImpute. Imputation was performed on aggregated representations (union aggregation, meaning feature values were merged using a logical OR operation; or average, meaning feature values were averaged).

Results Table 4 shows that SoftImpute achieves near-perfect performance on script vectors, with accuracy and F1 above 0.98 across all configurations. These results substantially exceed those for all types of typological vectors, indicating that script features are both more complete and predictable within genealogical lineages.

There are two factors that explain this result. First, script data are inherently less sparse (i.e. before imputation, script vectors cover 7,488 languages (nearly twice the most complete typological distance type, which is syntactic), and after imputation, coverage exceeds 20,000 languages). Secondly, writing system properties (e.g. the complexity of characters) are highly stable within language families (Miton and Morin, 2021), providing strong structural cues for imputation models.

Applying expanded lineage imputation improves most metrics across every vector type, with the largest gains for typological features (in particular morphological) under union aggregation. For instance, recall increases by 2.06% overall (5.24% for morphological) and F1 increases by 1.40% (3.10% for morphological), gains attributable to their initial sparsity. These results confirm that expanded lineage imputation improves imputation quality and is effective in reducing sparsity.

The exceptions are phonological and inventory features. Phonological features decline in accuracy (↓ 1.01%), precision (↓ 2.90%), and F1 (↓ 0.98%) following expanded lineage imputation, while inventory features decline in precision (↓ 0.27%). Both feature types also decline under average aggregation: phonological features worsen by 2.18% in RMSE and 0.77% in MAE, and inventory features by 0.94% in RMSE and 3.27% in MAE. This suggests that lineage propagation may not be well-suited for these feature types. Phonological features encode how sounds are produced, including manner and place of articulation. Meanwhile, inventory features capture the

Setup	POS	DEP	XNLI	Taxi1500	SIB200	POS2	DEP2	MT	EL
Baseline	18.5	73.0	67.6	23.3	8.38	33.8	36.3	35.9	72.1
1	**25.55** ↑38%	71.6 ↓1.9%	68.0 ↑0.6%	**25.51** ↑9.7%	8.84 ↑5.5%	34.1 ↑0.8%	38.7 ↑6.6%	31.7 ↓11.7%	75.4 ↑4.5%
2	19.0 ↑2.8%	73.0 ↓0.1%	70.0 ↑3.6%	**26.1** ↑12%	7.58 ↓9.5%	32.1 ↓5%	36.1 ↓0.5%	33.1 ↓7.6%	69.9 ↓3.1%
1 + 2	**27.03** ↑46%	72.3 ↓1%	67.2 ↓0.5%	**26.92** ↑15.7%	8.40 ↑0.2%	33.2 ↓1.7%	36.5 ↑0.6%	32.0 ↓10.8%	66.9 ↓7.2%
3	17.0 ↓8%	73.0	67.9 ↑0.4%	22.7 ↓2.4%	9.76 ↑16.5%	32.5 ↓3.7%	38.3 ↑5.5%	35.9 ↑0.1%	67.8 ↓5.9%
1 + 3	21.3 ↑14.8%	73.4 ↑0.5%	67.5 ↓0.1%	**25.33** ↑8.9%	9.01 ↑7.5%	34.0 ↑0.7%	37.6 ↑3.6%	33.4 ↓7%	69.7 ↓3.3%
2 + 3	13.1 ↓29.1%	74.5 ↑2.1%	71.5 ↑5.7%	23.5 ↑1.1%	9.87 ↑17.8%	36.3 ↑7.3%	39.1 ↑7.9%	34.9 ↓2.8%	61.5 ↓14.7%
1 + 2 + 3	18.6 ↑0.5%	72.6 ↓0.5%	70.9 ↑4.9%	25.0 ↑7.5%	9.79 ↑16.8%	36.3 ↑7.5%	38.0 ↑4.9%	**30.2** ↓15.8%	62.8 ↓12.9%

Table 6: NDCG@3 scores in LangRank tasks under our ablation study. Percentage changes are relative to the baseline. **Bolded** values indicate $p < 0.05$ in comparison to the baseline. Baseline is SoftImpute. 1 = Script; 2 = Glottolog integration; 3 = Lineage Impute.

phonetic sound inventories of languages. Neither set of properties may transfer consistently across related languages. Supporting this idea, Table 5 shows that phonological and inventory features also show the weakest performance of all vector types when evaluating lineage imputation quality alone, achieving the lowest scores across all metrics under both union and average aggregation.

Table 5 shows that lineage imputation performs strong across both typological and script vectors, with script vectors outperforming typological ones on every metric. This is likely because, as mentioned earlier, writing systems are more consistently inherited within language families, making parent-to-child propagation a particularly reliable assumption for script data.

Despite trailing script vectors, typological performance remains exceptionally high (F1 = 0.9790, Accuracy = 0.9886). The higher RMSE under average aggregation (0.1068 compared to 0.0366) suggests more continuous prediction error for typological features, which is expected since typological features have greater cross-linguistic variability.

These results confirm that our additions not only increase coverage but also improve the internal consistency and predictability of URIEL+, demonstrating the combined benefit of script vectors and expanded lineage imputation.

5.2. Benchmarking Cross-Lingual Transfer

Evaluation Setup Given the common application of language distances in transfer language selection (Blaschke et al., 2025; Rice et al., 2025; Ng et al., 2025, 2026), we evaluate the impact of our additions on cross-lingual transfer using the LangRank framework (Lin et al., 2019). LangRank, which ranks potential transfer languages for multilingual NLP tasks using gradient-boosted decision trees trained on language distances. Following the framework of Ng et al. (2026)[5], we apply LangRank to select transfer languages across nine representative sub-tasks: machine translation (MT), entity linking (EL), part-of-speech tagging (POS 1 & 2), dependency parsing (DEP 1 & 2), natural language inference (XNLI (Conneau et al., 2018)), and topic classification (Taxi1500 (Ma et al., 2025), SIB200 (Adelani et al., 2024)). This framework measures transfer language selection quality (i.e. how well language distances rank potential source languages for a given target language) and measures transfer language selection quality, specifically, how well language distances rank potential source languages for a given target language. The *Normalized Discounted Cumulative Gain (NDCG@3)* metric evaluates whether our improved distance calculations help identify the top-3 most promising source languages for training, compared to the empirically determined optimal source.

We conduct an ablation study to isolate the effect of each contribution and to quantify its practical impact on transfer predictions. Specifically, we train LangRank on URIEL+ distances following SoftImpute, under three configurations: (1) incorporating script distances, (2) integrating languages from Glottolog, and (3) conducting expanded lineage imputation prior to SoftImpute. Transfer Performance is evaluated using the Normalised Discounted Cumulative Gain at rank 3 (NDCG@3), which measures alignment between LangRank's predicted transfer language rankings and the optimal language rankings.task transferability.

Results Table 6 demonstrates LangRank performance across ablations. Overall, performance gains are task-specific, appearing incan be found in all but two tasks studied (MT, EL). In particular, incorporating script distances significantly improves LangRank transfer predictions in part-of-speech tagging and Taxi1500 topic classification. Simultaneously, expanded lineage imputation yields further improvements in SIB200 topic classification and DEP 2. Given the high language coverage in these tasks (799 in Taxi1500, 197 in SIB200 (Ng et al., 2026)), most of

[5]We employ the same LangRank hyperparameters and the same set of tasks.

which are low-resource, these improvements demonstrate how our extensions to low-resource language coverage support cross-lingual transfer.

Generally, while performance gains underscore the utility of our contributions in choosing transfer languages, improvements remain modest and fluctuate between tasks and setups. Simply incorporating script distance yields an average performance improvement of 6%, the highest among all other setups. Moreover, many improvements do not reach statistical significance, and our additions exhibit variable performance impacts across tasks and setups. However, our downstream evaluation has limited scope for assessing our additions. While we expand data coverage in URIEL+ for mostly low-resource languages, the downstream tasks primarily cover higher-resource languages already attested in URIEL+ previously. Thus, our downstream evaluation on transfer performance does not directly benchmark the expanded language coverage that we contribute. Therefore, we turn to task-level patterns to better characterize where our additions are effective.

The variation in performance gains across tasks is informative and it highlights the conditions under which our additions are most effective. For instance, script vectors significantly improve transfer language selection for POS tagging, where differences in writing system hamper transfer performance in morphological tasks (Murikinati et al., 2020), and where languages sharing a script tend to share tokenisation behaviour (Xhelili et al., 2024; Zhuang et al., 2025). Simultaneously, expanded lineage imputation yields gains in tasks which include a larger proportion of medium- and low-resource languages (SIB200, POS2 and DEP2), consistent with the reduced data sparsity of languages in those resource levels. Overall, our additions tend to improve transfer where they provide information relevant to the task. This finding is consistent with Ng et al. (2026), and underscores the importance of multi-faceted approaches to improving data coverage in URIEL+.

6. Conclusion

Our paper introduces a systematic approach to reducing sparsity in the URIEL+ linguistic knowledge base. By introducing script vectors for $7,488$ languages, incorporating $18,710$ languages from Glottolog, and extending lineage imputation to $26,449$ languages, we substantially improve data coverage, particularly for low-resource languages. Script vectors provide explicit representations of writing systems, supporting more reliable cross-lingual modelling. The Glottolog additions expand the knowledge base's genealogical breadth, while expanded lineage imputation propagates typological and script features through these genealogies, filling gaps for low-resource languages. These additions reduce feature sparsity, increase language coverage, and improve imputation quality metrics. Furthermore, we benchmark the additions applicability in cross-lingual transfer. By releasing our improvements, we provide the NLP community with a more complete and equitable foundation for multilingual research, particularly benefiting under-represented languages. We have provided a new release of URIEL+, with our additions found here: https://github.com/LeeLanguageLab/URIELPlus.

7. Future Work

The lineage imputation scheme we explore is generally motivated by the agreement in feature values between parent and child languages. However, it is rather simple. We plan to explore more advanced imputation schemes that take advantage of the tree-like structure of language phylogeny.

Furthermore, since URIEL+ is not synchronized with Glottolog and ScriptSource, we intend to periodically update it using new releases from these sources to maintain alignment with current language and script metadata.

8. Limitations

Our work relies on existing data sources such as ScriptSource and Glottolog. Therefore, it is subject to any biases or inaccuracies present in them. We could not perfectly replicate the baseline imputation quality results from URIEL+, despite running the same code provided by (Khan et al., 2025) and achieving the same feature coverage prior to imputation (see syntactic, phonological, inventory, and morphological feature coverage before expanded lineage imputation in Figure 2). Consequently, the distances used in our downstream experiments differed slightly, leading to minor variations in downstream performance.

Furthermore, the features themselves have inherent limitations. Script similarity does not always equate to the deep linguistic similarity required for complex downstream tasks. For example, script vectors correctly capture that Persian and Arabic share script features. However, Persian is an Indo-European language while Arabic is Semitic. For a task like machine translation, over-reliance on the script signal could incorrectly suggest a stronger linguistic relationship than exists, potentially adding a misleading signal.

Similarly, our expanded lineage imputation operates on the heuristic that dialects share features with their parent languages. While broadly effective, this assumption can be flawed. A dialect may have evolved a unique feature or, through language contact, lost a feature that its parent retained. In such cases, the imputation would be incorrect, adding a

faulty data point that could negatively impact downstream tasks. Furthermore, any errors or biases in the documented parent language features are systematically inherited by all descendants. This risks potentially amplifying inaccuracies in our dataset, particularly for parent language where prestige varieties are particularly favoured, as is common for many of the low-resource families our method targets (Joshi et al., 2020).

Finally, adding $18,710$ languages from Glottolog to URIEL+, each with genetic, geographic, typological (syntactic, phonological, inventory, and morphological), and script vectors, substantially increases the time required to integrate all linguistic data sources in URIEL+ (e.g. from a few minutes to often around an hour).

9. Ethics Statement

This work builds upon publicly available linguistic resources, including URIEL+, Glottolog, and ScriptSource. URIEL+ and Glottolog are both publicly viewable and accessible, whereas ScriptSource is publicly viewable but not openly accessible for bulk data extraction or redistribution. Access to ScriptSource data for this research was obtained with the explicit permission and assistance of the resource's administrator, to ensure compliant and responsible use. Furthermore, we received the permission of the URIEL+ authors to contribute and extend their work.

All data used in this work describe linguistic structures and writing systems, without involving personally identifiable or user-generated content. Our primary objective is to address data sparsity in multilingual linguistic knowledge bases by systematically enriching low-resource features rather than collecting new human data. While this enrichment improves coverage, automatic feature propagation through expanded lineage imputation may inadvertently reinforce existing typological biases from well-documented languages. To mitigate this, we ensured that the imputation procedure was transparent and reproducible and we release all derived resources and code under an open license to promote transparency and community oversight.

Acknowledgements

We thank Martin Raymond and ScriptSource for providing us the script feature data, Khasir Hean for his feedback on the expanded lineage imputation, and Patrick Littell for suggesting the use of ScriptSource for constructing our script vectors.

A. Additional Results for the Imputation Quality Test

In addition to the imputation quality tests done in Tables 4 and 5, another imputation quality test was done in Table 7. This test has values filled with the expanded lineage imputation procedure held out during imputation evaluation. The imputation results from this test are the results that users of URIEL+ would see when they integrate all databases and impute with both lineage imputation and SoftImpute and are the best imputation results of all quality tests done.

Table 7 shows that lineage imputation consistently improves SoftImpute performance across all vector types and aggregations. The gains are most pronounced for typological vectors, with recall increasing from 0.7092 to 0.8477 ($\uparrow$ 19.53%), F1 increasing from 0.7841 to 0.8938 ($\uparrow$ 13.99%), RMSE dropping from 0.2964 to 0.2175 ($\downarrow$ 26.68%), and MAE dropping from 0.1961 to 0.1282 ($\downarrow$ 34.69%), suggesting that lineage imputation meaningfully fills in feature gaps that SoftImpute alone struggles with. We can see even stronger individual improvements with recall and F1 for morphological vectors increasing from 0.5998 to 0.7223 ($\uparrow$ 20.42%) and 0.6962 to 0.7999 ($\uparrow$ 14.90%), respectively. We see RMSE and MAE for phonological vectors decreasing from 0.2570 to 0.1833 ($\downarrow$ 28.68%) and 0.1678 to 0.1087 ($\downarrow$ 35.22%), respectively. Script vectors show more modest but consistent improvements across all metrics, which is unsurprising given their already strong baseline performance.

The better performance when lineage imputation is considered alongside SoftImpute is because lineage imputation fills values that are highly predictable from family structure–cases where a child language inherits a feature directly from its parent. These are structurally easy cases, and SoftImpute tends to predict them well precisely because related languages cluster together in feature space. Including them in the evaluation therefore inflates scores relative to the harder, non-lineage-covered positions, which explains why the typological gains are so large. Typological features have broader lineage coverage, meaning more of these easy, family-consistent positions are added back into the evaluation pool.

B. Bibliographical References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Com-*

Stage	Vector	Union-Agg				Average-Agg	
		Accuracy	Precision	Recall	F1	RMSE	MAE
− Lineage	typ	0.8821	0.8767	0.7092	0.7841	0.2909	0.1914
	syn	0.8410	0.8496	0.6650	0.7461	0.3357	0.2467
	pho	0.9140	0.9454	0.8472	0.8936	0.2570	0.1678
	inv	0.9432	0.9267	0.8245	0.8726	0.2015	0.1040
	mor	0.8451	0.8295	0.5998	0.6962	0.3374	0.2473
	scr	0.9939	0.9949	0.9826	0.9887	0.0769	0.0248
+ Lineage	typ	0.9437 (↑ 6.98%)	0.9453 (↑ 7.82%)	0.8477 (↑ 19.53%)	0.8938 (↑ 13.99%)	0.2133 (↓ 26.68%)	0.1250 (↓ 34.69%)
	syn	0.8927 (↑ 6.15%)	0.9134 (↑ 7.51%)	0.7716 (↑ 16.03%)	0.8366 (↑ 12.13%)	0.2873 (↓ 14.42%)	0.2058 (↓ 16.58%)
	pho	0.9649 (↑ 5.57%)	0.9729 (↑ 2.91%)	0.9420 (↑ 11.19%)	0.9572 (↑ 7.12%)	0.1833 (↓ 28.68%)	0.1087 (↓ 35.22%)
	inv	0.9715 (↑ 3.00%)	0.9668 (↑ 4.33%)	0.9018 (↑ 9.38%)	0.9332 (↑ 6.94%)	0.1567 (↓ 22.23%)	0.0786 (↓ 24.42%)
	mor	0.8885 (↑ 5.14%)	0.8962 (↑ 8.04%)	0.7223 (↑ 20.42%)	0.7999 (↑ 14.90%)	0.2947 (↓ 12.66%)	0.2107 (↓ 14.80%)
	scr	**0.9949** (↑ 0.10%)	**0.9958** (↑ 0.09%)	**0.9851** (↑ 0.25%)	**0.9904** (↑ 0.17%)	**0.0729** (↓ 5.20%)	**0.0240** (↓ 3.23%)

Table 7: Performance of SoftImpute on typological and script vectors before (−) and after (+) expanded lineage imputation. Metrics are reported for union-aggregated and average-aggregated evaluation schemes. Values that are imputed through lineage imputation are included in the imputation evaluation. **Bolded** values indicate the best score for each metric.

putational Linguistics (Volume 1: Long Papers), pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2025. ProxyLM: Predicting language model performance on multilingual tasks via proxy models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1981–2011, Albuquerque, New Mexico. Association for Computational Linguistics.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Pro-*

ceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Olive Jean Dunn. 1961. Multiple Comparisons among Means. 56(293):52–64.

Gilles Guillot and François Rousset. 2013. Dismantling the mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344.

Nataliia Hübler. 2022. Phylogenetic signal and rate of evolutionary change in language structures. *Royal Society Open Science*, 9(3):211252.

Gerhard Jäger and Johannes Wahle. 2021. Phylogenetic Typology. *Frontiers in Psychology*, 12.

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, David Anugraha, Jinyu Liu, Jiaxu Li, 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.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

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. 2019. 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.

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, and Hinrich Schuetze. 2025. LangSAMP: Language-script aware multilingual pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1743–1770, Vienna, Austria. Association for Computational Linguistics.

Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. Taxi1500: A dataset for multilingual text classification in 1500 languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.

Nathan Mantel. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2):209–220.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research: JMLR*, 11:2287–2322.

Helena Miton and Olivier Morin. 2021. Graphic complexity in writing systems. *Cognition*, 214:104771.

George Moroz. 2017. *lingtypology: easy mapping for Linguistic Typology*.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

York Hay Ng, Phuong Hanh Hoang, and En-Shiun Annie Lee. 2025. Less is more: The effectiveness of compact typological language representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25805–25816, Suzhou, China. Association for Computational Linguistics.

York Hay Ng, Aditya Khan, Xiang Lu, Matteo Salloum, Michael Zhou, Phuong H. Hoang, A. Seza Doğruöz, and En-Shiun Annie Lee. 2026. Modality matching matters: Calibrating language distances for cross-lingual transfer in uriel+.

Vít Novotný, Eniafe Festus Ayetiran, Dalibor Bačovský, Dávid Lupták, Michal Štefánik, and Petr Sojka. 2021. One size does not fit all: Finding the optimal subword sizes for FastText models across languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1068–1074, Held Online. INCOMA Ltd.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

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.

Wondimagegnhue Tufa, Ilia Markov, and Piek Vossen. 2024. Unknown script: Impact of script on cross-lingual transfer. 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 124–129, 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.

Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. 2015. Multi-level block permutation. *NeuroImage*, 123:253–268.

Orgest Xhelili, Yihong Liu, and Hinrich Schuetze. 2024. Breaking the script barrier in multilingual pre-trained language models with transliteration-based post-training alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11283–11296, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2025. Enhancing cross-lingual transfer through reversible transliteration: A Huffman-based approach for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16299–16313, Vienna, Austria. Association for Computational Linguistics.

C. Language Resource References

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

Harald Hammarström and Robert Forkel. 2022. *Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information*.

Holloway, Steph. 2013. *ScriptSource - Writing systems, computers and people*. SIL International.

Khan, Aditya and Shipton, Mason and Anugraha, David and Duan, Kaiyao and Hoang, Phuong H. and Khiu, Eric and Doğruöz, A. Seza and Lee, En-Shiun Annie. 2025. *URIEL+: Enhancing Linguistic Inclusion and Usability in a Typological and Multilingual Knowledge Base*. Association for Computational Linguistics.

Koopman, Hilda. 2009. *Syntactic Structures of the World's Languages (SSWL)*.

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

Littell, Patrick and Mortensen, David R. and Lin, Ke and Kairis, Katherine and Turner, Carlisle and Levin, Lori. 2017. *URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors*. Association for Computational Linguistics.

Marsico, Egidio and Flavier, Sebastien and Verkerk, Annemarie and Moran, Steven. 2018. *BDPROTO: A Database of Phonological Inventories from Ancient and Reconstructed Languages*. European Language Resources Association (ELRA).

Skirgård, Hedvig and Haynie, Hannah J and Blasi, Damián E and Hammarström, Harald and Collins, Jeremy and Latarche, Jay J and Lesage, Jakob and Weber, Tobias and Witzlack-Makarevich, Alena and Passmore, Sam and others. 2023. *Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss*. American Association for the Advancement of Science.