Modality Matching Matters: Calibrating Language Distances for Cross-Lingual Transfer in URIEL+

Anonymous ACL submission

Abstract

Existing linguistic knowledge bases such as URIEL+ provide valuable geographic, genetic and typological distances for cross-lingual transfer but suffer from two key limitations. One, their one-size-fits-all vector representations are ill-suited to the diverse structures of linguistic data, and two, they lack a principled method for aggregating these signals into a single, comprehensive score. In this paper, we address these gaps by introducing a framework for type-matched language distances. We propose novel, structure-aware representations for each distance type: speaker-weighted distributions for geography, hyperbolic embeddings for genealogy, and a latent variables model for typology. We unify these signals into a robust, task-agnostic composite distance. In selecting transfer languages, our representations and composite distances consistently improve performance across a wide range of NLP tasks, providing a more principled and effective toolkit for multilingual research.

1 Introduction

Linguistic knowledge bases such as URIEL/URIEL+ (Littell et al., 2017; Khan et al., 2025) are foundational tools that quantify linguistic distance for over $7,000$ languages. These distances fall into three *modalities*, or feature categories: geographic (locations of languages), genetic (linguistic family trees), and typological (linguistic features unique to each language)[1], as shown in Figure 1. These measures are widely used in cross-lingual transfer research to assess and leverage linguistic similarity between languages for tasks such as selecting optimal source languages for model training (Lin et al., 2019; Lauscher et al., 2020; Ruder et al., 2021; Khiu et al., 2024).

As indicated by Toossi et al. (2024), URIEL represents languages in all three modalities as high-dimensional Euclidean vectors, compared via angular distance. This uniform approach is convenient but ill-suited for the diverse structures of linguistic data. That is to say, it produces less meaningful distances and limits the effectiveness of cross-lingual transfer where accurate representations of linguistic distance is paramount. In our study, we address this issue by proposing modality-specific distances from new language representations.

Limitations in URIEL+ Representations

Geographic Both URIEL and URIEL+ represent each language by a single Glottolog coordinate, with geographic vectors computed as great-circle distances to 299 fixed reference points. This single-point proxy misses multi-country and diaspora populations. It also reflects historical or administrative geographical locations rather than current speaker distributions which is a key determinant for language contact (Nichols, 1992). For example, English, French, and Spanish are pinned near cities such as London, Paris, and Madrid, although most speakers of these languages reside elsewhere (Figure 1, Geographic). This can result in counter-intuitive discrepancies, causing languages with large, overlapping speaker communities to appear geographically distant and providing misleading signals for transfer.

Genetic The current genetic representation flattens the Glottolog tree into sparse, one-hot vectors indicating language family membership (>3700 dimensions, 99.85% zeros), losing the crucial hierarchical structure of genetic relationships. This flat representation counts shared ancestry at all levels equally. For example, the close relationship between German and English (West Germanic) is given the same weight as the far more distant relationship between German and Hindi (Indo-European) (Figure 1, Genetic), obscuring fine-grained distinctions essential for transfer.

[1]The typological modality is also commonly referred to as featural (e.g. in Khan et al., 2025).

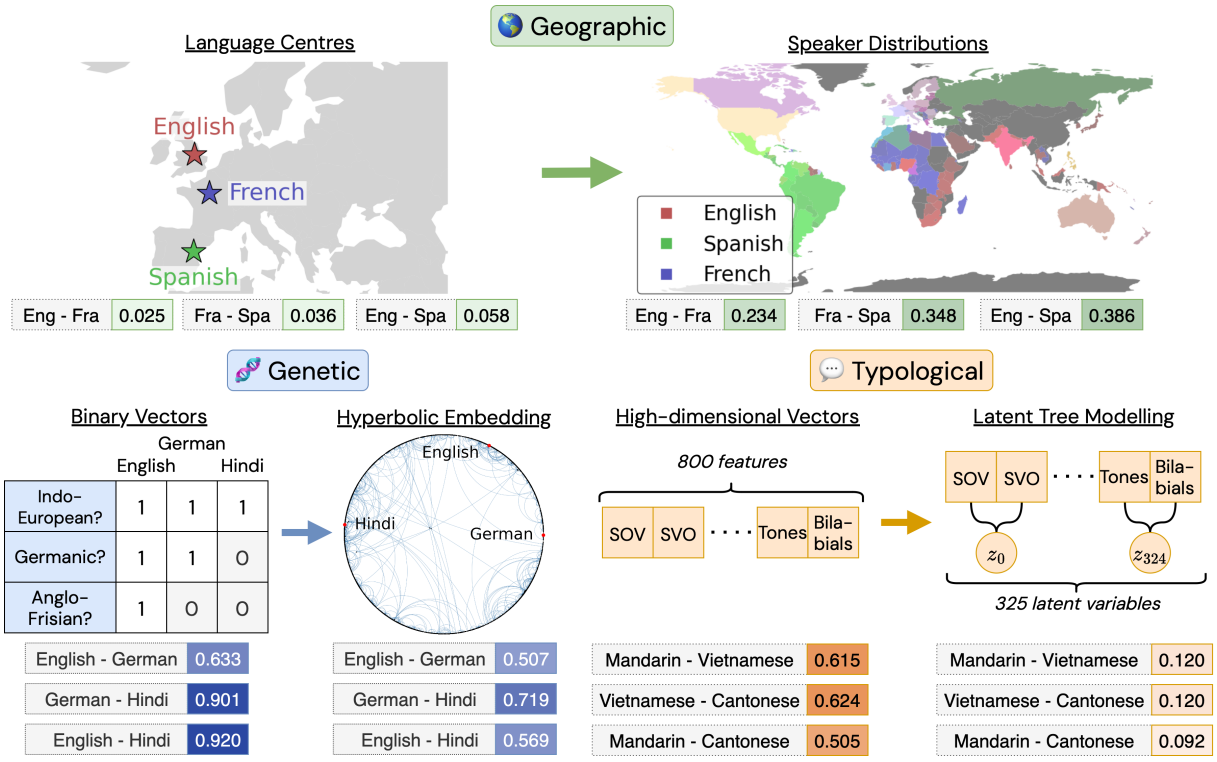

Figure 1: A demonstration of URIEL+ language representations versus our proposed representations, for each modality. Distance scores are shown for URIEL+ (left number) and our proposed representation (right number). Lower values indicate greater similarity.

Moreover, this representation is limited to terminal nodes (languages), failing to provide embeddings for internal nodes (language families and sub-families), which are crucial for historical linguistics.

Typological High-dimensional binary feature vectors are sparse, with correlated and sometimes redundant features, weakening the ability of angular distances to capture meaningful structural similarity. For instance, features for "Subject-Object-Verb" and "Subject-Verb-Object" word order are highly correlated yet treated as independent signals, inflating distances between languages which differ on related features. Ng et al. (2025) showed that such redundancy and high dimensionality reduce the effectiveness of typological vectors in capturing meaningful structural similarity.

Given the limitations in language representations in URIEL and URIEL+ (especially for cross-lingual transfer), what makes a good language distance for transfer? We claim that each modality should use a representation and distance suited to its structure. Therefore, we embed the original URIEL+ vectors into a form that captures the inherent structure (e.g., the hierarchical genealogy) of each modality and compute distances on this new representation.

Another fundamental limitation of URIEL+ is that it cannot compute a cumulative distance using all modalities. This forces researchers to choose between signals (e.g., typology or genetics), even though a unified metric is often preferred for practical applications such as transfer language selection (Ahuja et al., 2022; Srinivasan et al., 2021). We address this gap by developing a composite distance: a weighted average of distances from individual modalities, providing a single value that simplifies applications in cross-lingual transfer.

Our paper rectifies the aforementioned issues with the following contributions:

1. We formalize modality-matched language distances, introducing new representations and distance metrics for each modality.

2. We propose a simple composite distance that aggregates modality-specific distances.

Empirically, across the LANGRANK setting and other transfer benchmarks, modality-matched distances consistently improve source language selection.

2 Related Research

URIEL in Cross-Lingual Transfer URIEL distances serve as a strong predictor of transfer performance (Khiu et al., 2024; Philippy et al., 2023; Lauscher et al., 2020; Tran and Bisazza, 2019) between languages, performing comparably to other linguistic measures (Eronen et al., 2023).

Consequently, URIEL distances have been widely applied to enhance cross-lingual transfer, particularly in predicting the performance of multilingual models (Anugraha et al., 2025; Srinivasan et al., 2021; Xia et al., 2020; Patankar et al., 2022), selecting transfer languages (Lin et al., 2019; Eronen et al., 2023), and language model regularization (Adilazuarda et al., 2024), demonstrating its indispensable role in multilingual natural language processing (NLP).

Distributional Representation of Geographic Data Moving from "language as a point" to "language as a distribution" is crucial for capturing signals from language contact (Dunn and Edwards-Brown, 2024; Nichols, 1992). Empirical audits show that single-point geography can mask biases in data by under-representing where speakers actually reside (Faisal et al., 2022). A natural method for comparing speaker distributions is the Wasserstein-1 distance (or Earth Mover's distance) (Villani, 2009), which measures the minimum "work" needed to transform one distribution into another. Optimal transport has proven effective in NLP for tasks such as measuring document similarity (Kusner et al., 2015), evaluating text generation (Clark et al., 2019), and aligning word embeddings (Zhang et al., 2017), making it a well-grounded choice for our geographic modality.

Sparser Representations of Typological Data Typological feature sets are often high-dimensional, redundant, and noisy (Ng et al., 2025), with inconsistent feature choices yielding wide variation across studies (Ploeger et al., 2024; Poelman et al., 2024). Compact, structured representations can mitigate these issues, enhancing the utility of typological distances in downstream NLP tasks such as machine translation and evaluation set selection (Ng et al., 2025; Bjerva, 2024; Ramiz and Shahbaz, 2025; Ploeger et al., 2025; Hlavnova and Ruder, 2023).

To achieve this, we turn to latent tree models (LTMs), which can uncover hidden structure from data without supervision. By grouping correlated features and capturing unobserved confounders, LTMs produce task-agnostic, denoised embeddings (Zwiernik, 2017; Williams et al., 2018) that have proven effective for related tasks such as topic discovery and sentence modeling (Mourad et al., 2013; Chen et al., 2016; Williams et al., 2018).

Hyperbolic Representations of Genetic Data Euclidean space (with flat curvature and polynomial volume growth) poorly fits data where latent structure is hierarchical or tree-like, and leads to unnecessary distortion. URIEL+ vectors lie such a flat space (see Appendix C). Instead, hyperbolic geometry offers a closer match as its exponential volume growth aligns with the branching of trees, enabling low-distortion, low-dimensional embeddings. Nickel and Kiela (2017) showed that Poincaré-ball embeddings capture WordNet hierarchies with markedly less distortion and in fewer dimensions than Euclidean baselines. Extending this idea, Tifrea et al. (2018) adapted the commonly used GloVe model to learn directly in hyperbolic space, improving word similarity, analogy, and especially hypernymy detection. Beyond the Poincaré model, the hyperboloid (Lorentz) model embeds points in Minkowski space, simplifying certain operations and often improving numerical stability during training (Nickel and Kiela, 2018).

In multilingual NLP, incorporating linguistic genealogy assists cross-lingual transfer (e.g., by guiding meta-learning with genetic structure or by arranging adapter modules to mirror the language tree (Garcia et al., 2021; Faisal and Anastasopoulos, 2022)). Prior hyperbolic work on languages used cognate similarity to infer hierarchical relations (Nickel and Kiela, 2018). To the best of our knowledge, our work is the first to directly embed the comprehensive language hierarchy from Glottolog (Hammarström et al., 2025) to hyperbolic space, providing a novel application and a rigorous empirical comparison of foundational geometric embedding techniques on this linguistic resource.

Need for a Composite Distance Score A recurring challenge in cross-lingual work is the need to juggle multiple, often task-dependent, linguistic distances without a single, reusable score. While resources such as Khan et al. (2025) provide individual distances, they do not offer a principled way to aggregate them. Some methods fuse modalities within a training objective (e.g., LINGUALCHEMY regularises with typological, geographic, and ge-

Modality	$\mathcal{X}^m$	$\mathcal{Z}^m$	f^m	d^m
Geography	Country L1 speaker counts + centroids	Distribution over locations (speaker shares)	Normalise counts to a probability distribution	Earth Mover's distance
Genetic	Glottolog genealogy	Hyperboloid Embeddings	Learn embeddings	Hyperbolic distance
Typology	Binary features	Posteriors over latent "islands"	Fit islands; map to posteriors	Angular distance

Table 1: Summary of modality representations and their distances. Distances are normalised and may be aggregated into a composite distance.

netic vectors), but these do not yield a calibrated, standalone language-to-language distance metric (Adilazuarda et al., 2024). This motivates our goal of creating a single, normalized composite score usable across tasks and languages.

Representation Requirements From Prior Work Synthesizing the evidence above, we adopt four requirements for cross-lingual distance:

- **Geography as distributions**: Languages should be represented as dispersed speaker distributions, not as single points.
- **Genealogy as hierarchy**: Distances should respect language ancestor–descendant structure.
- **Typology as low-noise factors**: Redundant/correlated features should be compressed into a compact representation.
- **Composability**: Modality-specific distances should be normalised so they can be aggregated into a single composite score.

3 Modality Representations and Cross-Modal Composition

The central premise in this work is that each modality benefits from a representation that matches its latent structure. To illustrate, we briefly review the modalities in URIEL+, and introduce our modality matched representations and distances along with describing we may combine them. A summary of the representations is presented in Table 1.

3.1 Formalizing Modalities

Let $\mathcal{L}$ denote the set of languages and let M denote modalities in URIEL+:

$$M = \{\text{geography},\ \text{genetic},\ \text{typology}\}.$$

For each modality $m \in M$, let $\mathcal{X}^m$ be the raw data space (e.g. country/territory speaker counts for geography, the Glottolog genealogy counts for genetic, derived latent vectors from typological information). For a language $\ell \in \mathcal{L}$, we write $x_\ell(m) \in \mathcal{X}^m$ for its raw modality-specific data. For example, $x_{\text{German}}(\text{geo})$ corresponds to the geography vector for the German language in URIEL+.

For each $m \in M$ we specify a representation mapping $f^m : \mathcal{X}^m \to \mathcal{Z}^m$, where $\mathcal{Z}^m$ is an appropriate representation space. For instance, if $m =$ genetic, then $\mathcal{Z}^m$ has to capture the hierarchical structure of the family tree of a particular language. After representing each modality vector for a language ℓ in the new representation space, denoted $f^m(x_\ell(m))$, we compute distances between these using a normalised distance $d^m \in [0, 1]$ defined on $\mathcal{Z}^m$.

3.2 Geography as Distributions

Representing a language with a single point ignores effects from language contact, arising from multi-country speaker populations shaped by globalization and migration. Modeling languages by the geographical distribution of speakers captures dispersion and overlap across regions, providing a population-aware geographic signal that better reflects the geographic proximity of languages.

We source from Ethnologue (Eberhard et al., 2025) the number of L1 language speakers per language per country to model each language as a discrete probability distribution over locations, with mass proportional to the share of speakers at those locations. We use L1 speaker counts from Ethnologue due to its broad language coverage and standardized data collection. However, we acknowledge that this presents reproducibility challenges (see the limitations section). In particular, for language $\ell \in \mathcal{L}$, let the location (i.e. countries or territories) where ℓ is spoken be indexed by $i = 1, \ldots, m$, with geographic centroids $y_i \in \mathbb{S}^2$ (WGS84) and L1 speaker counts $n_{\ell,i} \geq 0$ (Karney, 2013). To calculate the distance between these

speaker distributions, we normalize speaker counts $n_{\ell,i}$ in each location i, yielding the share of speakers of language ℓ at location i, $w_{\ell,i}$. This produces the distribution $\mathbb{P}_\ell = \{(y_i, w_{\ell,i})\}_{i=1}^m$. Essentially, each language ℓ is represented by a list of locations (represented as coordinates) with weight $w_{\ell,i}$ corresponding to the proportion of L1 speakers residing there. We, therefore, define f^{geo} as the mapping $x_\ell(\text{geography}) \mapsto \mathbb{P}_\ell$.

A natural distance measure d^{geo} between speaker distributions is the Earth Mover distance. To define it, suppose that $\ell_1 \mapsto \mathbb{P}_{\ell_1} = \{(y_i, w_{\ell,i})\}_{i=1}^m$ and $\ell_2 \mapsto \mathbb{P}_{\ell_2} = \{(z_i, v_{\ell,i})\}_{i=1}^n$. We define the set of feasible transport plans

$$\Pi(\mathbb{P}_{\ell_1}, \mathbb{P}_{\ell_2}) = \left\{ \pi \in \mathbb{R}_{\geq 0}^{m \times n} \;\middle|\; \begin{smallmatrix} \sum_j \pi_{ij} = w_i \\ \sum_i \pi_{ij} = v_j \end{smallmatrix} \right\}$$

Allowing us to define language distance as

$$d^{\text{geo}}(\ell_1, \ell_2) = \frac{1}{D_{\max}} \min_{\pi \in \Pi} \sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij} d_g(y_i, z_j)$$

where d_g is the shortest distance between the two geographic centroids that remain on the Earth's surface, also known as the geodesic distance; and $D_{\max} = \max_{x,y \in \mathbb{S}^2} d_g(x, y)$, representing the geodesic distance between the two poles on Earth. This metric iterates through all possible methods of transforming one speaker distribution into another, choosing the one requiring the least work. After normalization, this yields a distance between speaker distributions. A proof that this normalization is valid is provided in Appendix B.

3.3 Genealogy as Hierarchy

To overcome the issues described in section 1, we propose a principled, structure-preserving approach by learning dense embedding vectors for the entire Glottolog genealogical tree, including families, languages, and dialects, in a low-dimensional, continuous space. The ideal geometric space for this task is hyperbolic geometry, whose metric properties are intrinsically suited for representing hierarchical data with minimal distortion. The space's negative curvature and exponential volume growth provide a natural geometric analogue to the branching, tree-like structure of linguistic evolution, where the number of descendants grows exponentially with depth from the proto-language root. Formally, we represent the Glottolog genealogy tree as a directed acyclic graph $G = (V, E)$, where V is the set of linguistic entities (nodes), and E contains the directed parent-to-child edges. Our goal is to learn an embedding function $f^{\text{gen}} : V \to \mathcal{H}^d$ that maps each node $v \in V$ to a point in the d-dimensional hyperbolic space. We explored two isometric models of hyperbolic geometry: the Poincaré disk model and the hyperboloid model, and denote the hyperbolic distance between a and b as $d_{\text{Hyp}}(a, b)$. The learning objective is designed to encourage the geometric arrangement of embeddings in $\mathcal{H}^d$ to faithfully reflect the complete genealogical topology of G. To enforce this globally, we define our set of positive training pairs, $\mathcal{P}$, as the transitive closure of the parent-child edges in E, meaning that a pair $(u, v) \in \mathcal{P}$ if and only if u is an ancestor of v. Hence, following (Nickel and Kiela, 2017, 2018), for each positive pair $(u, v) \in \mathcal{P}$, we adopt a contrastive objective, sampling K negative nodes $\{w_1, \ldots, w_K\}$ that are not descendants of u, and define the objective per pair as

$$\mathrm{L}_{(u,v)} = \frac{\exp(-d(u, v))}{\sum_{k=1}^{K} \exp(-d(u, w_i))}.$$

The total objective is $\mathrm{L}_{(u,v)}$ summed over all positive pairs: $\mathrm{L} = \sum_{(x,y) \in \mathcal{P}} \mathrm{L}_{(x,y)}$. Maximizing this objective pulls each positive pair closer to each other while simultaneously pushing negative pairs farther apart, thus encouraging hierarchical fidelity.

The derived distance metric on $\mathcal{Z}^m$ is given by $d^{\text{gen}} = d_{\text{Hyp}}(a, b)/D_{\max}$. Here $D_{\max}$ is the maximum pairwise hyperbolic distance. This ensures that the distance is bounded in $[0, 1]$. In preliminary experiments, the hyperboloid model performed stronger in ancestor retrieval tasks. Thus, we adopt the hyperboloid embeddings and distance metric for LANGRANK experiments and evaluation. Further details are in Appendix C.

3.4 Typology as Low-Noise Factors

A natural choice to understand confounding variables and inherent structure in language typology is latent tree models (LTM). We use this to cluster typological features into groups (termed "islands") governed by latent variables that capture confounding variables, co-occurence structure, while addressing redundancy. We obtain a dimensionality reduction mapping f^{typ} from this method.

Given a subset of binary typological features $w_\ell = (w_{\ell,1}, \ldots, w_{\ell,m})$, we introduce a latent variable Z with $k = 2$ states and conditional probabilities $\theta_{jk} := \mathbb{P}(w_{\ell,j} = 1 \mid Z = k)$ learned by

Table 2: List of the NLP tasks applied to LANGRANK. "Target" and "Source" refers to the number of source and target languages where models are tested and trained on, respectively. Related works link to previous applications in choosing transfer languages based on langauge distances.

Task Type	Dataset	Related Work	Model	Metric	Target	Source
Machine Trans.	TED	Lin et al. (2019)	RNN+Attn	BLEU	54	54
Dep. Parsing	UD v2.2	Lin et al. (2019)	Biaffine	Accuracy	30	30
	UD v2.14	Blaschke et al. (2025)	UDPipe 2	LAS	152	70
POS Tagging	UD v2.2	Lin et al. (2019)	BiLSTM	Accuracy	60	26
	UD v2.14	Blaschke et al. (2025)	UDPipe 2	UPOS	152	70
Entity Linking	Wikipedia	Lin et al. (2019)	BiLSTM	Accuracy	54	9
Topic Class.	Taxi1500	–	mBERT	Macro F1	799	33
	SIB200	Blaschke et al. (2025)	XLM-R	Macro F1	197	160
NLI	XNLI	Philippy et al. (2023)	mBERT	Accuracy	15	15

Expectation–Maximization (Dempster et al., 1977), where priors are initialized uniformly and conditionals are initialized randomly. We perform early stopping via a modified Bayesian Information Criterion (BIC) [2] which penalizes log-likelihood and the number of parameters quadratically, encouraging more balanced clusters.

To scale beyond a single latent variable, we implement a greedy algorithm to obtain multiple "islands". Iteratively, we repeat the following process: (i) initialize an active set using the pair of features with highest Mutual Information (MI) (Peng et al., 2005) not yet assigned to any latent variable; (ii) add the feature yielding the highest MI with the features in the active set; (iii) attempt to split the active set into two using the modified BIC; (iv) if the split is preferred, refine by testing feature switches across the two groups to further improve BIC. When a split is accepted, we obtain two groups G_1, G_2. We define the larger group as an island, associating it with a latent variable z_i, and store its $m_i \times 2$ parameter matrix (θ_{jk}) as a cluster. The remaining features return to the pool and the process repeats.

Finally, a typological vector $x_\ell(\text{typ})$ is mapped to the concatenated posterior vector

$$\mathbf{p}(w_\ell) := \big(\mathbb{P}(z_i = 0 \mid w_\ell),\ \mathbb{P}(z_i = 1 \mid w_\ell)\big)_{i=1}^{n\ \top}.$$

where n is the number of islands. This representation is naturally normalized per island. We compute angular distances on our representation, as is done by default in Khan et al. (2025), due to its sensitivity to the proportional relationships between posterior probabilities across islands, rather than their absolute magnitudes; thus making it a robust metric for comparing the structural profiles of languages.

[2] See Appendix D for implementation details.

3.5 Composability: Aggregating Distances

Practitioners often desire a single distance score between languages. Given nonnegative modality weights $w \in \mathbb{R}_{\geq 0}^{|M|}$ with $\sum_{m \in M} w_m = 1$, we define the normalised composite distance

$$D(\ell_i, \ell_j) := \sum_{m \in M} w_m d^m(f^m(x_{\ell_i}(m), x_{\ell_j}(m)).$$

Although the weights can be learned specifically for a given cross-lingual transfer task, the simplest case is to simply let $w_m = 1/|M|$ for all m. In doing so, D collapses to a simple average–this serves as a strong default. It assumes the user does not want to favour any particular modality when evaluating how distant language is. Furthermore, it is simple and robust, requiring no task-specific tuning. Nonetheless, we present alternative ways to select weights in Appendix E.

4 Validation on Downstream Tasks

Although prior work on evaluating distance measures have mostly explored the impact of individual distances on transfer performance (Lauscher et al., 2020; Philippy et al., 2023; Blaschke et al., 2025), we aim to faithfully illustrate the utility of our language representations in enhancing cross-lingual transfer by applying LANGRANK (Lin et al., 2019), a widely used framework for choosing transfer (source) languages for cross-lingual NLP tasks utilizing decision trees.

Table 3: The impact of distance metrics on performance loss when picking the top transfer language from LangRank. Values are regression coefficients ± standard error, measured in percentage points. Baseline rows represent the intercept, indicating the performance loss when using URIEL+ representations for each modality. Lower is better. Results where $p < 0.05$ are shown in **bold**. Color corresponds to the percentage change in performance loss.

Modality	Representation	DEP	EL	MT	POS
Baseline:		11.4 ± 2.9	30.0 ± 6.2	12.5 ± 1.8	27.9 ± 4.4
Typ	Laplacian	0.8 ± 1.0	-3.8 ± 2.8	0.7 ± 0.9	-2.1 ± 1.9
	Islands	0.5 ± 1.0	-1.2 ± 2.8	-1.0 ± 0.9	-0.4 ± 1.9
Geo	Speaker	0.6 ± 0.7	$\mathbf{-7.4 \pm 2.0}$	-1.0 ± 0.6	-0.3 ± 1.3
Gen	Hyperbolic	-0.9 ± 0.7	3.6 ± 2.0	$\mathbf{-4.5 \pm 0.6}$	-1.0 ± 1.3

Modality	Representation	Taxi1500	SIB200	XNLI	UD2.8 POS	UD2.8 DEP
Baseline:		38.1 ± 0.5	16.9 ± 1.1	6.2 ± 1.2	27.4 ± 1.5	35.6 ± 1.9
Typ	Laplacian	0.4 ± 0.3	-0.2 ± 0.5	0.4 ± 0.6	$\mathbf{1.8 \pm 0.8}$	1.5 ± 0.9
	Islands	$\mathbf{-0.9 \pm 0.3}$	$\mathbf{-1.4 \pm 0.5}$	$\mathbf{-2.4 \pm 0.6}$	-0.6 ± 0.8	$\mathbf{-1.8 \pm 0.9}$
Geo	Speaker	$\mathbf{-2.1 \pm 0.2}$	-0.6 ± 0.3	0.1 ± 0.4	$\mathbf{-1.6 \pm 0.6}$	0.7 ± 0.6
Gen	Hyperbolic	$\mathbf{2.7 \pm 0.2}$	$\mathbf{1.0 \pm 0.3}$	-0.1 ± 0.4	$\mathbf{-2.6 \pm 0.6}$	$\mathbf{-3.9 \pm 0.6}$

4.1 Experimental Setup

Table 2 lists the tasks studied. Based on the findings in Blaschke et al. (2025), we augment the original framework with five new tasks: Taxi1500 (Ma et al., 2025), due to its sizeable language coverage; XNLI (Conneau et al., 2018), SIB200 (Adelani et al., 2024), along with dependency parsing and part-of-speech tagging tasks from Universal Dependencies (Nivre et al., 2020), where the relationship between transfer performance and language distance was previously determined (Philippy et al., 2023; Blaschke et al., 2025). This expanded evaluation serves to support the generalizability of our findings across tasks and languages.

We utilize "performance loss" to measure how well LANGRANK enhances cross-lingual performance in NLP tasks. Performance loss is defined as the relative loss in performance when transferring from the top-1 language chosen by LANGRANK, compared to the performance of the optimal source, for a given target language. [3] This setup demonstrates the real-world impact of language representations on cross-lingual transfer more accurately.

Using only language distances as features, we conduct an ablation study by training LANGRANK with distances from different representations [4]. For the genetic modality, we ablate on the URIEL+ and hyperbolic representations; for the typological modality, we additionally ablate on the representation applying Laplacian Score feature selection (He et al., 2005) on URIEL+ typological vectors, which was found to be a robust selection method for LANGRANK in Ng et al. (2025). Within each ablation and task, we conduct leave-one-language-out cross-validation (i.e. testing performance loss for each target language).

Collecting scores across folds and ablations, we fit a linear mixed-effects model with performance loss as the dependent variable, three categorical variables indicating the representation used as fixed effects, and a random intercept measuring baseline URIEL+ performance. Model parameters are estimated via L-BFGS optimization. This approach allows us to estimate the impact of each representation, while accounting for variability across folds.

4.2 Results

The impact of our new representations on cross-lingual transfer performance is detailed in Table 3. First, we observe that baseline performance losses varied from 6.2 - 38.1 between tasks, confirming that, even when applying URIEL+ distance measures, LANGRANK remains a viable and robust choice for choosing transfer languages.

Next, there usually exists some combinations of language representations that significantly improve cross-lingual performance. Notably, our representations can substantially reduce transfer error. For example, in the XNLI task, using our latent islands representation for typology reduces the baseline performance loss of 6.2 by 2.4 points (a 39% improvement). Similarly, for Machine Translation, our hyperbolic genetic embeddings reduce

[3] See Appendix F.2 for the formal definition.
[4] See Appendix F for the full setup, and hyperparameters.

the baseline loss of 12.5 by 4.5 points (a 36% improvement).

These consistent reductions in performance loss highlight how our representations generally outperform URIEL+, in particular for the low-resource languages in our evaluation (e.g. Taxi1500 contains 764 low-resource languages[5]). Through aligning representations and distance metrics with the inherent structure of linguistic modalities, our framework unlocks more nuanced signals for cross-lingual transfer.

These results however illustrate a cautionary tale: although our representations can significantly improve performance, there are instances where swapping out URIEL+ representations worsens performance. This task-dependent variability suggests a deeper interplay between the nature of a task and the linguistic information most salient to it. We hypothesize, for instance, that tasks highly sensitive to language contact and lexical borrowing, such as certain classification or entity linking tasks, benefit most from our speaker distribution model, which explicitly captures geographic overlap.

Conversely, tasks focused on deep syntactic structure might have a more complex relationship with genealogy; while our hyperbolic embeddings faithfully model the Glottolog hierarchy, the transferability of syntax may be influenced more by recent, horizontal contact phenomena or areal features not captured by vertical descent alone. The finding that the impacts of each new representation are task-dependent aligns with Blaschke et al. (2025); therefore, in there exists no one-size-fits-all solution for cross-lingual transfer.

Table 4: Performance loss when choosing top-1 transfer languages using composite distances. Lower is better. Compare with scores in the baseline row in Table 3.

Task	DEP	EL	MT	POS	XNLI
Score	9.9	25.6	11.2	22.8	3.5
Task	Taxi	SIB	POS 2	DEP 2	
Score	46.7	14.4	21.3	36.7	

Composite Distances. We additionally benchmark the performance loss incurred when choosing transfer languages based on the composite distance measure from Section 3.5. Defining $w_m = \frac{1}{|M|}$, this distance measure averages over distances from our new representations in each modality.

The utility of this composite distance is shown in Table 4. Our results demonstrate that the composite distance serves as a strong general-purpose baseline. On most of the tasks evaluated, including Entity Linking (25.6 vs. a baseline of 30.0) and XNLI (3.5 vs. a baseline of 6.2), it reduces performance loss compared to using URIEL+ distances alone. However, its substantial under-performance on tasks such as Taxi1500 classification (46.7 loss vs. a baseline of 38.1) highlights that a simple, unweighted average can obscure the most important modality for certain applications, reinforcing the need for our flexible, task-aware toolkit.

This metric addresses a long-standing need in the community for a single, reliable score for language similarity. Additionally, our framework enables future work in learning weights based on relevance to specific tasks, yielding supplementary performance gains while gaining insights into the linguistic aspects most salient for different tasks.

5 Conclusion

We presented a new framework for computing linguistic distance based on modality-matched representations. Our novel, structure-aware methods for geography (speaker distributions), genealogy (hyperbolic embeddings), and typology (latent feature islands) were designed to better capture the unique characteristics of each linguistic signal.

Our experiments confirm that the utility of these representations is fundamentally task-dependent–no single metric is optimal for all scenarios. This finding reframes our contribution as a flexible toolkit for cross-lingual research, empowering practitioners to choose the most suitable distance metric for their specific application. As a general alternative, we proposed a composite distance that averages these signals. While this score provides a strong, general-purpose baseline that improves over URIEL+ on a majority of the tasks we tested, its under-performance on some tasks highlights that a simple, unweighted average can obscure the most important modality for certain applications, reinforcing the need for our flexible, task-aware toolkit. To encourage community participation, we release all our code for more principled investigations into linguistic distance: https://anonymous.4open.science/r/language-representations.

[5] Applying the definition of "low-resource language" from Joshi et al. (2020).

Limitations

Data Sources. Our work fundamentally relies on existing linguistics sources, and therefore inherits any inaccuracies or incomplete data, which may affect the quality of language representations unequally. In particular:

- Our speaker distribution model is founded on the basis that geographic proximity of speakers influence language contact, but this model is constrained by the granularity and scope of Ethnologue. It relies on national-level speaker counts, which may not accurately capture the precise distribution of speakers. Additionally, Ethnologue does not consider other factors influencing speaker interactions, such as time, topography, and culture. Furthermore, as the data from Ethnologue is proprietary, this prevents us from publicly releasing our representations.

- Hyperbolic embeddings are designed to solely model the Glottolog tree. However, Glottolog represents only one specific model of language history that is subject to ongoing linguistic research and revision.

- Our latent feature islands method offer another representation of URIEL+'s typolpogical data, but remains subject to the issue of sparsity. Specifically, 87% of values in URIEL+ are missing (Ng et al., 2025). This impacts the accuracy of our representations, with potentially more pronounced effects on low-resource languages.

Evaluation Scope. Our evaluation was conducted across a diverse but limited set of NLP tasks. Since the effects of language representations have been shown to be task-specific, the proposed representations are not guaranteed to be applicable to other tasks not studied here. Our results further demonstrate variability in performance even within the same tasks (such as between XNLI and SIB200), likely originating from other factors such as data domain, choice of model, language coverage, etc. Moreover, we focus on the application of language distances on choosing transfer languages using LANGRANK only; the utility of our language representations on other frameworks and/or applications remains unexplored.

Distance Measures. While our work demonstrates the strength of distances from new language representations, these singular numerical distances, even in a focused direction, cannot fully capture the complexity in linguistic relationships. Furthermore, the task-agnostic composite distance we present should not be considered as universally effective. More complex, non-linear models, adapted to specific tasks, could potentially yield further gains, which we leave for future work.

To mitigate these issues and promote accessibility, we release our full codebase. Furthermore, while the speaker distributions cannot be released due to data licensing, we publicly release our Hyperbolic genetic embeddings and Latent Island typological representations to encourage more principled investigations into linguistic distance.

Ethics Statement

The intention of this study is to enhance the representations of the world's languages, with the ultimate aim of improving cross-lingual performance, while promoting equity and inclusivity, in language technologies.

No personally identifiable or sensitive data was used in this study. However, our work relies on established linguistic knowledge bases and datasets, and we acknowledge that our work is subject to any biases or inaccuracies in these sources, which may particularly under-represent low-resource languages or certain speaker communities.

We further recognize that our proposed methods may be computationally intensive, which can create barriers for researchers with limited computational resources. To promote accessibility and reproducibility, we release our code and language representation data where possible.

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

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Ayu Purwarianti, and Alham Fikri Aji. 2024. LinguAlchemy: Fusing ty-

pological and geographical elements for unseen language generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3912–3928, Miami, Florida, USA. Association for Computational Linguistics.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. 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. 2024. The role of typological feature prediction in nlp and linguistics. *Computational Linguistics*, 50(2):781–794.

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.

Peixian Chen, Nevin L. Zhang, Tengfei Liu, Leonard K. M. Poon, Zhourong Chen, and Farhan Khawar. 2016. Latent tree models for hierarchical topic detection.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn and Lane Edwards-Brown. 2024. Geographically-informed language identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7672–7682, Torino, Italia. ELRA and ICCL.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Ethnologue: Languages of the world. twenty-eighth edition.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Jezabel Garcia, Federica Freddi, Jamie McGowan, Tim Nieradzik, Feng-Ting Liao, Ye Tian, Da-shan Shiu, and Alberto Bernacchia. 2021. Cross-lingual transfer with MAML on trees. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 72–79, Kyiv, Ukraine. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2. Accessed: 2025-09-16.

Xiaofei He, Deng Cai, and Partha Niyogi. 2005. Laplacian score for feature selection. *Advances in neural information processing systems*, 18.

Ester Hlavnova and Sebastian Ruder. 2023. Empowering cross-lingual behavioral testing of NLP models with typological features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198, Toronto, Canada. Association for Computational Linguistics.

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.

Charles F. F. Karney. 2013. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. 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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

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.

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.

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.

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.

R. Mourad, C. Sinoquet, N. L. Zhang, T. Liu, and P. Leray. 2013. A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47:157–203.

York Hay Ng, Phuong Hanh Hoang, and En-Shiun Annie Lee. 2025. Less is more: The effectiveness of compact typological language representations.

Johanna Nichols. 1992. *Linguistic diversity in space and time*. University of Chicago Press.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "typological diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2025. A principled framework for evaluating on typologically diverse languages.

Wessel Poelman, Esther Ploeger, Miryam de Lhoneux, and Johannes Bjerva. 2024. A call for consistency in reporting typological diversity. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 75–77, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Ramiz and Muhammad Shahbaz. 2025. The use of linguistic typology in understanding language universals.

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.

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

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincaré glove: Hyperbolic word embeddings.

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.

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Cédric Villani. 2009. *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg.

Adina Williams, Andrew Drozdov, and Samuel R. Bowman. 2018. Do latent tree learning models identify meaningful structure in sentences?

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.

Daniel Zeman et al. 2024. Universal dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Piotr Zwiernik. 2017. Latent tree models.

A Language Coverage of Representations

We report the number of languages covered by our language representations in Table 5.

Representation	Number of languages
Speaker	6695
Hyperbolic	7836
Islands	4555

Table 5: Number of languages with data per representation.

Although URIEL+ nominally enables distance computations for 8171 languages, coverage within each modality varies, as the underlying data sources contain information for only subsets of languages. Our proposed representations are subject to similar limitations, namely being constrained by the language coverage of Ethnologue, Glottolog, and URIEL+. The hyperbolic embeddings represent families, languages, and dialects, totaling 26223 entities. Moreover, the combined breadth of these resources remains considerable, underscoring their utility in cross-lingual transfer particularly for less-resourced languages.

B Geographic Distance Metric Derivations

Here, we prove the normalization property of the geographic distance we discuss in section 3.2.

Denote the Wasserstein-1 distance by W_1. We know that for any two languages P, Q we have $W_1(P, Q) \leq D_{\max}$ because we can always design a transport plan π such that

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} c(x_i, y_j) \leq D_{\max}.$$

The details of this plan π are as follows. For every (i, j) pairing, we set $\pi_{ij} = w_i \cdot v_j$. We first check that this is a valid transport plan.

1. It is clear that for all i, j, $w_i, v_j \geq 0$, $\pi_{ij} \geq 0$.
2. For any i, we see that $\sum_{j=1}^{n} \pi_{ij} = \sum_{j=1}^{n} (w_i \cdot v_j) = w_i \sum_{j=1}^{n} v_j = w_i \cdot 1 = w_i$.
3. For any j, we see that $\sum_{i=1}^{m} \pi_{ij} = \sum_{i=1}^{m} (v_j \cdot w_i) = v_j \sum_{i=1}^{m} w_i = v_j \cdot 1 = v_j$.

Hence, this is a valid plan. Then, we know that for any two points on earth x, y, that $d_g(x, y) = c(x, y) \leq D_{\max}$. Therefore, plugging this inequality into the above summation using the aforementioned transport plan gives us that

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} c(x_i, y_j) \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} D_{\max} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} (w_i \cdot v_j) D_{\max} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} (w_i \cdot v_j) D_{\max} \\
&= D_{\max} \sum_{i=1}^{m} w_i \sum_{j=1}^{n} v_j \\
&= D_{\max}
\end{aligned}$$

Now, from the definition of Wasserstein-1 distance, we know that

$$W_1(P, Q) \leq \sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} c(x_i, y_j) \leq D_{\max},$$

and this statement is proved. In addition, normalizing based on antipodal distance is also the technique implemented by URIEL+, which gives credence to this normalization technique.

C Genetic Embedding: Geometry & Optimization Details

This appendix collects the implementation details that were omitted from the main body but are necessary to reproduce the genetic embeddings in each geometry.

C.1 Poincaré Ball Model

We work in the open unit ball $\mathcal{B}^d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < 1\}$ endowed with the Riemannian metric

$$g_{\mathbf{x}} = \left(\frac{2}{1 - \|\mathbf{x}\|_2^2}\right)^2 I_d.$$

Translations use Möbius addition

$$\mathbf{u} \oplus \mathbf{v} = \frac{(1 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|_2^2)\mathbf{u} + (1 - \|\mathbf{u}\|_2^2)\mathbf{v}}{1 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2},$$

with the denominator clamped to $\geq \epsilon$. The optimization uses Riemannian stochastic gradient descent. Given a Euclidean gradient g_e, it is first converted to a Riemannian gradient in the tangent space of $\mathbf{x}$ by scaling:

$$g_r = \frac{(1 - \|\mathbf{x}\|_2^2)^2}{4} g_e.$$

The update is then performed by moving along the geodesic in the direction of $-g_r$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t \oplus \left(\tanh\left(\frac{\eta \lambda_{\mathbf{x}_t} \|g_r\|_2}{2}\right) \frac{-g_r}{\|g_r\|_2}\right),$$

where η is the learning rate. After the update, if a point $\mathbf{y}$ lands outside the unit ball due to numerical instability, it is projected back to the boundary by rescaling: $\mathbf{y} \leftarrow \mathbf{y} \frac{1-\epsilon}{\|\mathbf{y}\|_2}$. For the geodesic distance (defined in the main body), the argument of $\cosh^{-1}(\cdot)$ is clamped to $\geq 1 + \epsilon$ for numerical stability.

C.2 Hyperboloid Model

We embed in

$$\mathcal{H}^d = \{\mathbf{x} \in \mathbb{R}^{d+1} : \langle \mathbf{x}, \mathbf{x} \rangle_L = -1, x_0 > 0\}$$

with Lorentzian inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_L = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i.$$

For the hyperbolic distance (defined in the main body), we clamp $-\langle \mathbf{u}, \mathbf{v} \rangle_L$ to $\geq 1 + \epsilon$. Optimization in the hyperboloid model is performed by applying the following update steps for a point $\mathbf{x}$ with a corresponding Euclidean gradient g_e:

1. Gradient Projection: The Euclidean gradient g_e is projected onto the tangent space at $\mathbf{x}$ to obtain the Riemannian gradient g_r. Let g_e^L be the gradient with its time-like coordinate negated. Then,
$$g_r = g_e^L + \langle \mathbf{x}, g_e^L \rangle_L \mathbf{x}.$$
2. Gradient Clipping: The norm of the Riemannian gradient is clipped to a maximum value of c_g:
$$g_r \leftarrow g_r \cdot \min\left(1, \frac{c_g}{\|g_r\|_L}\right).$$
3. Exponential Map: The point is updated by moving along the geodesic. The tangent vector for the update is $\mathbf{u} = -\eta g_r$, where η is the learning rate. This produces an intermediate point, $\tilde{\mathbf{x}}$:
$$\tilde{\mathbf{x}} = \cosh(\|\mathbf{u}\|_L)\mathbf{x}_t + \sinh(\|\mathbf{u}\|_L)\frac{\mathbf{u}}{\|\mathbf{u}\|_L}.$$
4. Manifold Projection: As a final safeguard, the intermediate point $\tilde{\mathbf{x}}$ is projected back to the hyperboloid to yield the final updated point $\mathbf{x}_{t+1}$. This step also prevents numerical overflow by clipping the norm of the spatial components of $\tilde{\mathbf{x}}$ (denoted $\tilde{\mathbf{x}}_{1:}$) to a maximum of c_s:
$$\mathbf{x}_{t+1} = \left[\sqrt{\|\mathbf{x}'_{1:}\|_2^2 + 1}, \mathbf{x}'_{1:}\right]$$
$$\text{where } \mathbf{x}'_{1:} = \tilde{\mathbf{x}}_{1:} \cdot \min\left(1, \frac{c_s}{\|\tilde{\mathbf{x}}_{1:}\|_2}\right).$$

The clipping thresholds c_g and c_s are hyperparameters.

C.3 Reconstruction Metrics and Results

To evaluate how well the learned embeddings capture the original hierarchical structure, we perform a link prediction task focused on ancestor-descendant relationships. For each node u in the graph V, we rank all other nodes $v \in V \setminus \{u\}$ based on their geometric distance $d(u, v)$ in ascending order. We treat the set of true ancestors of u, denoted $\mathcal{A}(u)$, as the positive items to be retrieved. From this ranking, we compute two retrieval metrics: Mean Rank (MR) and Mean Average Precision (MAP).

Mean Rank (MR) This metric measures the average rank of a true ancestor. For each descendant-ancestor pair (u, a) where $a \in \mathcal{A}(u)$, we compute the rank of a in the distance-sorted list of nodes relative to u. A lower MR indicates better performance, as it means true ancestors are, on average, found closer to their descendants in the embedding space. The rank is formally defined as: $\text{rank}(a, u) = 1 + |\{v \in V \setminus (\mathcal{A}(u) \cup \{u\}) : d(u, v) < d(u, a)\}|$. The final MR is the average of these ranks over all true descendant-ancestor pairs in the graph.

Mean Average Precision (MAP) MAP provides a more comprehensive measure of ranking quality by rewarding models that place many true ancestors early in the ranked list. For each node u, we first compute its Average Precision (AP), which is the average of precision values at each rank k that contains a true ancestor:
$$\text{AP}(u) = \frac{\sum_{k=1}^{|V|-1} P(k) \times \mathbb{I}(v_k \in \mathcal{A}(u))}{|\mathcal{A}(u)|},$$
where v_k is the node at rank k, $P(k)$ is the precision at rank k (i.e., the fraction of true ancestors in the top k results), and $\mathbb{I}(\cdot)$ is the indicator function. The final MAP score is the mean of these AP scores over all nodes in the graph. A higher MAP score indicates better performance.

Results The performance of our genetic embedding algorithm across different geometries and dimensions is summarized in Table 6. The results clearly show that hyperbolic geometries (Hyperboloid and Poincaré) significantly outperform Euclidean geometry, especially at lower dimensions. The Hyperboloid model consistently achieves the best scores, demonstrating its effectiveness in capturing the hierarchical relationships of the data. Hence, we select the Hyperboloid model.

D Implementation Details for Latent Tree Models.

We employ a modified Bayesian Information Criterion (BIC) defined as $2k^2 log(n) - 2l$, where k denotes the number of parameters, l is the log-likelihood, and n is the number of samples. This modified criterion, which penalizes the number of parameters quadratically, more strongly discourages models with a large number of free parameters compared to the traditional linear penalty. In our greedy clustering context, this helps prevent the

Table 6: Reconstruction performance on the ancestor retrieval task. We report Mean Rank (MR) and Mean Average Precision (MAP) for each geometry across varying embedding dimensions (Dim).

Geometry	Dim	MR	MAP
Hyperboloid	2	6.3329	0.6743
	5	2.5227	0.8723
	10	1.3674	0.9513
	50	1.2518	0.9581
Poincaré	2	6.9936	0.5969
	5	2.1246	0.8601
	10	2.0591	0.8633
	50	2.1478	0.8463
Euclidean	2	274.0730	0.1910
	5	147.7106	0.3043
	10	56.3716	0.4286
	50	3.3975	0.7180

algorithm from forming many small, fragmented clusters, instead favoring more balanced and structurally coherent feature islands. This modified criterion tends to produce more balanced clusters than the conventional BIC formulation. When computing the BIC values for two clusters, there is a higher penalty for having imbalanced cluster sizes.

To learn a latent variable for a subset of features, we run the Expectation–Maximization algorithm with five restarts with random initializations to mitigate the risk of convergence to local optima.

The resulting model yields 325 feature clusters, each associated with a latent variable. Cluster sizes range from 1 to 11. To assess whether the algorithm effectively groups correlated features, we compute the absolute Pearson correlation among features within each cluster as a measure of intra-cluster association strength. For clusters of size three or larger, the average absolute correlation is 0.623, indicating that features grouped together tend to be strongly correlated. Clusters of size one or two are excluded from this analysis, as their small size suggests that those features are not substantially correlated with others.

E Task-Specific Weights For Composite Distances

Although one can learn the weights in a number of different ways, we present one simple method using the performance losses from our LANGRANK evaluation framework. If $l \in [0, 1]$ is a performance loss (e.g. accuracy, F1, or RMSE if it is known to be in the unit interval), then $1-l$ gives a measure of the quality of performance on a given task. In this case, one can use each of the modality distances d^m as covariates to predict l, say via a linear regression. Upon obtaining the coefficient estimates, one can take the coefficients into $[0, 1]$. Common options include transforming each coefficient estimate by the logistic function (or ReLU) and then normalizing.

F Downstream Task Setup Details

Our objective is to design an evaluation (tasks, evaluation metric) which is closely aligned with actual applications of language distances in cross-lingual transfer. In particular, the usage of language distances on choosing source languages has been widely studied (see Section 2). We therefore focus on applying new language representations to LANGRANK (Lin et al., 2019), a commonly used framework for choosing source languages for a given NLP task.

We mostly replicate Lin et al. (2019) and Khan et al. (2025)'s pipeline for evaluating distances using LANGRANK. This process involves first collecting, for a given NLP task (e.g. Taxi1500 topic classification) and model (e.g. mBERT), a dataset of performance scores for each target and source language pair. Next, during evaluation, we perform leave-one-language-out cross-validation by holding out scores for each target language, training a LightGBM ranker on the remaining data (additionally holding out 10% of data as a validation set), and evaluating the ranker on how well it picks source languages for the held-out target language.

F.1 Experimental Datasets

With LANGRANK, we evaluate the utility of distances by applying them to a diverse set of nine sub-tasks. For the first four (DEP, EL, MT, POS) we re-use the performance datasets provided by Lin et al. (2019). We additionally derived performance datasets for each new task studied:

- **Taxi1500:** Due to the infeasibility of training models for each language covered by Taxi1500, we train 33 mBERT (Devlin et al., 2019) models according to the languages in Taxi1500 which are defined as high- or medium-resource in URIEL+, evaluating each model's performance on the 799 languages

whose data is publicly available and contains >900 examples.

- **SIB200 & XNLI:** We train one model for each language, (in SIB200, rejecting 37 languages where the model did not converge), and finally evaluating each model on the test splits of all other languages.

- **UD v2.14:** We replicate the setup from Blaschke et al. (2025), and simply evaluate the test split of each language on each of the 70 UDPipe2 (Straka, 2018) models, averaging scores over treebanks within the same language.

For each task, we use the same train-validation-test splits as published.

F.2 Evaluating Distances

After collecting datasets, we run LANGRANK and ablate on, for each modality, training with distances computed from the URIEL+ representation versus our new representation. We measure its performance with the performance loss metric l, which are averaged across folds, to better showcase the real-world implications of our LANGRANK experiments. Here, we define performance loss l_i for the fold associated with holding out target language i as:

$$l_i = \frac{(\max_j s_{ij}) - s_{ij}}{\max_j s_{ij}}$$

where j is the top-1 language chosen by LANGRANK, and score s_{ij} refers to the model performance on the given NLP task when transferring to language i from language j. Simply put, given a particular model and a particular NLP task, performance loss l measures the relative difference in model performance between transferring using LANGRANK's chosen language and the optimal language.

In particular, we choose to consider only the top-1 chosen language due to the observation that practitioners often choose only the top-1 language (as opposed to, e.g. trying all top-3 languages) to perform cross-lingual transfer. This decision therefore aligns with our underlying objective of designing a realistic evaluation setup.

F.3 Computational Setup

Hyperparameters. We adopt the following hyperparameters for the LightGBM ranker:

- Early stopping rounds: 25
- Learning rate: 0.1
- Min data in leaf: 10
- Lambda L2: 0.2

These hyperparameters were obtained upon performing a grid search, and measuring the task-averaged LANGRANK performance when using baseline URIEL+ distances.

For training transfer models in tasks Taxi1500 and SIB200, since per-language data is relatively scarce ($\sim$1k examples), we employ the following training arguments:

- Num train epochs: 10
- Learning rate: 1e-5
- Batch size: 16
- Eval steps: 20
- Early stopping patience: 5
- Weight decay: 0.01
- Warmup ratio: 0.1

For XNLI, we replicate the setup from Philippy et al. (2023), with the following training arguments:

- Num train epochs: 3
- Learning rate: 2e-5
- Batch size: 32

Computing Infrastrcture. Model training and evaluation for collecting LANGRANK experimental datasets were conducted on a single NVidia A100, requiring around 100 compute hours.

All actual LANGRANK experiments were performed on an Apple M1 Pro over 8 hours.

G Licenses for Artifacts Used

The artifacts employed in this study, along with their respective licenses, are listed in Table 7.

All artifacts and datasets were used for the purpose of studying language representations, and were handled in accordance with their respective licenses.

Artifact	License
Packages	
URIEL+ (Khan et al., 2025)	CC BY-SA 4.0
LANGRANK (Lin et al., 2019)	BSD 3-Clause
Datasets	
Glottolog (v5.2) (Ma et al., 2025)	CC BY 4.0
Ethnologue (Edition 28) (Eberhard et al., 2025)	Proprietary (Licensed under SIL International)
Taxi1500 (v3) (Ma et al., 2025)	Apache 2.0
XNLI (Conneau et al., 2018)	CC BY-NC 4.0
SIB200 (Adelani et al., 2024)	CC BY-SA 4.0
UD (v2.14) (Zeman et al., 2024)	Various
Models	
Multilingual BERT cased (Devlin et al., 2019)	Apache 2.0
XLM-RoBERTa-base (Conneau et al., 2018)	CC BY-NC 4.0
UDPipe v2.12 (Straka, 2018)	MPL 2.0

Table 7: Artifacts used in this study, and their licenses.

H Use of Generative AI

Generative AI was employed only in a limited capacity: to assist in organizing and clarifying text, and to suggest code auto-completions during the implementation of experiments.