

Size Sound Symbolism Survives Genealogy Across 27 Languages

Anika Sharma¹ Emma Wrenn² Tianyi Niu³ Shashank Srivastava¹

¹University of North Carolina at Chapel Hill

²McGill University ³Capital One

Abstract

Sound symbolism, the non-arbitrary mapping between word sounds and meanings, is well known from effects like Bouba–Kiki. But its cross-linguistic scope is difficult to assess because phonological similarity can also arise from shared ancestry, contact, or language-specific inventories. We propose a genealogy-aware stress test for size sound symbolism across 27 languages spanning 13 language families, using 810 phonemically transcribed size adjectives validated with native-speaker audio. We treat cross-family iconicity as a hard generalization task: train on languages grouped by lexical distance to a held-out target, test transfer from increasingly distant languages, and apply adversarial pressure against similarity-bin and source-language cues. Interpretable bag-of-phoneme classifiers predict size above chance, with clear patterns involving /i/ for smallness and /o,a/ for largeness, while consonants add more language-specific cues. Adversarial training prevents the jointly trained model from exploiting genealogy while held-out size prediction stays above chance, though post-hoc probing shows language identity is still recoverable. The residual signal is small but consistent, indicating cross-family size cues.¹

1 Introduction

If you were watching a superhero movie called *Lamonians vs. Grataks: The Phoneme Accords*, chances are you would already be rooting for the Lamonians because they *sound* nicer. In a 2009 *Guardian* article, linguist David Crystal noted that people tend to favor the fictional Lamonians, drawn to the soft consonants (/l/, /m/, /n/) and long vowels that give the name a gentler tone (Crystal, 2009). Such intuitions reflect *sound symbolism*: systematic associations between sound and meaning that

challenge the classical assumption that linguistic form is arbitrary. But do these intuitions reveal broad cognitive regularities, or merely shared ancestry, contact, and borrowing?

Sound symbolism is familiar in onomatopoeia, as in *buzz* or *crash*, but it also appears in less direct forms across languages: Yucatec Maya uses vowel length to signal event duration, while Japanese food mimetics use consonantal contrasts to convey crispness (Raeveskiy et al., 2023). Establishing cross-linguistic sound symbolism, however, is difficult. Apparent universals may reflect genealogy or areal contact, and languages differ in which sounds they contain. Any genuine cross-family effect is therefore likely subtle; large effects are more plausibly genealogical or contact-based. Beyond theoretical linguistics, this question matters for perceptual grounding, low-resource cross-lingual transfer, and applications such as brand naming.

We treat cross-family sound symbolism as a hard generalization problem, training on some languages and testing on a held-out language to see what persists as the training set becomes less related. We focus on size adjectives and ask: **When size adjectives are tested across increasingly distant languages, and models are discouraged from using source-language cues, do phonological size cues still survive?** To answer this, we build an evidence ladder: distance-stratified leave-one-language-out transfer, feature analyses, ablations, and adversarial training against genealogy/source-language cues. Our contributions are:

- A cross-linguistic dataset of 800+ size adjectives from 27 languages across 13 families, phonemically transcribed in IPA and validated with native-speaker recordings.
- A genealogy-aware adversarial stress test for size symbolism: size prediction remains above chance under both similarity-bin adversarial training (54.4%) and stricter 26-way source-

¹Code, data, and model checkpoints for the experiments are available at https://github.com/Anika1703/testing_size_symbolism.

language adversarial training (55.8%), while adversarial heads fall near chance. Post-hoc probes show source-language identity remains recoverable, so we interpret these results as testing cue usability rather than full language removal.

- Feature analyses showing a vowel-dominant cross-family pattern (/i/ for smallness and /o,a/ for largeness), with consonants contributing additional, more language-conditioned cues.

Our aim is not to maximize classification accuracy, but to test whether size remains detectable after genealogical shortcuts are pressured away.

2 Related Work

The classical linguistic view treats the form–meaning link as largely *arbitrary* (Saussure, 2011), with exceptions in *iconicity* and *sound symbolism*. Foundational studies show systematic mappings between phonological form and perceived magnitude: Sapir (1929) found reliable size judgments for nonce words differing only in vowel quality, while others showed cross-cultural mappings between sound and shape (Köhler, 1929; Ćwiek et al., 2022). Mechanistic accounts often link lower-frequency sounds, such as back/low vowels, to *largeness*, and higher-frequency sounds to *smallness* (Ohala, 1994). At the same time, surveys emphasize that iconic effects are typically *modest* and *domain-specific* (Johansson and Zlatev, 2013).

Magnitude symbolism: Size symbolism is one of the best-studied domains. Prior work associates high front vowels such as /i/ with “smallness” and back or low vowels such as /o, a/ with “largeness” (Brown et al., 1955; Jespersen, 1922; Newman, 1933; Ultan, 1978; Berlin, 1994), consistent with acoustic explanations based on formants and the frequency code (Ohala, 1994). Consonantal effects have also been observed, including affricates, trills, and place/manner contrasts, as in North American diminutives (Nichols, 1971) and links between trilled /r/ and tactile “roughness” (Winter et al., 2022). However, such effects may be lexically constrained: in English, some size-related effects in adjectives do not generalize to other parts of speech (Winter and Perlman, 2021).

Other iconic domains and cross-family patterns: Iconic mappings extend beyond magnitude to shape, texture, motion, affect, and social stance. Boubá–kiki patterns appear across languages (Ćwiek et al., 2022; Sidhu et al., 2021; Passi

and Arun, 2024), and related effects have been reported for speed, brightness, and sentiment (Dingemanse et al., 2016; Sidhu and Pexman, 2019). Sound-symbolic systems also occur across many families, including Japanese mimetics (Hamano, 1986), Dravidian and Austroasiatic expressives (Emeneau, 1995; Annamalai, 1968; Diffloth, 1972, 1976), African ideophones (Childs, 1994, 2008; Samarin, 1965), and English morphophonological patterns (Bolinger, 1940; Rhodes and Lawler, 1981; Calvillo-Torres et al., 2024). Yet apparent universals are difficult to interpret: shared ancestry and contact can mimic cross-linguistic regularities.

Computational approaches and genealogy control:

Computational studies of form–meaning associations across languages find small but reliable biases (Blasi et al., 2016; Gutiérrez et al., 2016; Pimentel et al., 2019, 2021; Winter et al., 2022; Kilpatrick et al., 2023). Recent work also shows that multimodal and language models can recover iconic patterns from limited phonetic input (Alper and Averbuch-Elor, 2023; Loakman et al., 2024). However, three gaps remain. (1) *Out-of-family validation*: many studies do not require generalization to less-related languages. (2) *Consonant coverage*: consonants receive less systematic analysis than vowels. (3) *Genealogical confounds*: prior work often uses stratification or mixed-effects modeling (e.g., Blasi et al., 2016; Pimentel et al., 2021), but rarely tests whether size-predictive structure remains when source-language cues are made difficult to exploit. We address these gaps by using lexical/topological distances (LDN) to form minimally related partitions (Petroni and Serva, 2010; Bakker et al., 2009) and pairing stratified evaluation with adversarial representation learning (Elazar and Goldberg, 2018; Liu et al., 2020).

Scope of our work: We build a genealogy-aware stress test for size sound symbolism. We:

1. Construct a curated, *audio-validated*, multi-family dataset of size *adjectives*;
2. Test *out-of-family* generalization by training within LDN/LDND similarity tertiles and evaluating on held-out languages;
3. Analyze both *vowels and consonants*, separating recurring vowel patterns from more language-conditioned consonantal cues;
4. Use an *adversarial probe* to test whether size prediction persists when similarity-bin and source-language cues are harder to exploit;
5. Run targeted phonetic ablations to identify

which cues support cross-family transfer.

3 Method

In this section, we describe the dataset, similarity-binning protocol, baseline classifiers, and adversarial models used to test whether size-predictive phonological cues persist when genealogy- and source-language cues are harder to exploit.

3.1 Data Collection

Language selection and word-list compilation.

We compiled a dataset of 810 size adjectives across 27 languages spanning 13 genealogical groupings, including 16 Indo-European and 11 non-Indo-European languages. Each language contains 30 items: 15 denoting smallness (e.g., *tiny*, *minuscule*) and 15 denoting largeness (e.g., *huge*, *enormous*). Languages were selected using the World Atlas of Language Structures (Dryer and Haspelmath, 2013) to maximize genetic and areal diversity while remaining feasible for native-speaker recruitment.

Adjective lists were constructed by translating English seed words using DeepL (DeepL SE, 2025), Google Translate (Google, 2025), and bilingual dictionaries (Wehr and Cowan, 1994; Clines and Stec, 2018; Kotey, 1998). We manually filtered borrowings and transliterations to prioritize native lexical items. Words were phonemically transcribed in IPA. Because pronunciations can vary by speaker and region, our final transcriptions reflect the native-speaker recordings we collected and have been checked by a trained phonetician. Since our models use bag-of-phoneme features rather than acoustic measurements, the analysis abstracts away from fine allophonic variation. We treat such variation as a limitation of our analyses, rather than assuming it is eliminated. Full word lists and transcriptions are provided in Appendix B.

Phonemic transcription pipeline. To ensure transcription quality, we used a hybrid pipeline. Where available, we first generated IPA forms with the XPF corpus, a rule-based phoneme-mapping resource (Priva et al., 2021). We then collected native-speaker audio recordings of the adjective lists through language-exchange platforms such as Tandem (Tripod Technology GmbH, 2025). A trained phonetician reviewed the recordings with acoustic analysis tools and manually revised or reconstructed each transcription. When no initial transcription was available, forms were created directly from the recordings. For several languages,

domain experts such as language instructors or researchers also reviewed the word lists or IPA annotations. Appendix H reports details of phoneme coverage by language.

Genealogical distance. To control for shared ancestry, we used a precomputed matrix of normalized Levenshtein distances (LDN; Appendix A) compiled by Nielsen (2017) following Bakker et al. (2009). These distances average phonological edit distance over 40-word Swadesh lists. For each held-out target language, we rank the remaining languages by LDN and divide them into equal-count similarity tertiles, keeping the resulting training sets comparable in size. Because these bins are defined by rank, they should be interpreted as target-relative rather than as fixed absolute-distance ranges. Their absolute separation therefore varies across targets. For example, for Italian, a Romance language with several close relatives in the sample, mean LDN increases from 0.72 in the Most Similar bin to 0.92 in the Least Similar bin. By contrast, for Hungarian, one of six relatively isolated targets, it increases only from 0.89 to 0.93. Additionally, in our study, LDN captures lexical relatedness, but should not be assumed to also capture similarity in phoneme inventories. To check this distinction, we computed inventory distances using PanPhon (Mortensen et al., 2016), which represents phonemes through articulatory features, for all 351 unique language pairs. These distances showed only a weak association with LDN (Pearson $r = .255$; Spearman $\rho = .211$; both $p < .0001$). We therefore treat genealogical distance and phoneme-inventory similarity as distinct sources of cross-linguistic similarity rather than interpreting the LDN bins as controlling for both. Furthermore, we use LDN rather than the further-normalized LDND variant following Petroni and Serva (2010), who argue that LDND can remove genealogical signal relevant to language classification. As a robustness check, we compare LDN and LDND rankings in our sample. The rankings are strongly correlated, and all target-relative tertile assignments are identical under the two metrics; we report the LDN/LDND formulas and rank-correlation analysis in Appendix C.

3.2 Baseline Classifiers

We first train interpretable classifiers to predict whether a word denotes smallness or largeness. Each word is represented as a bag-of-phoneme vec-

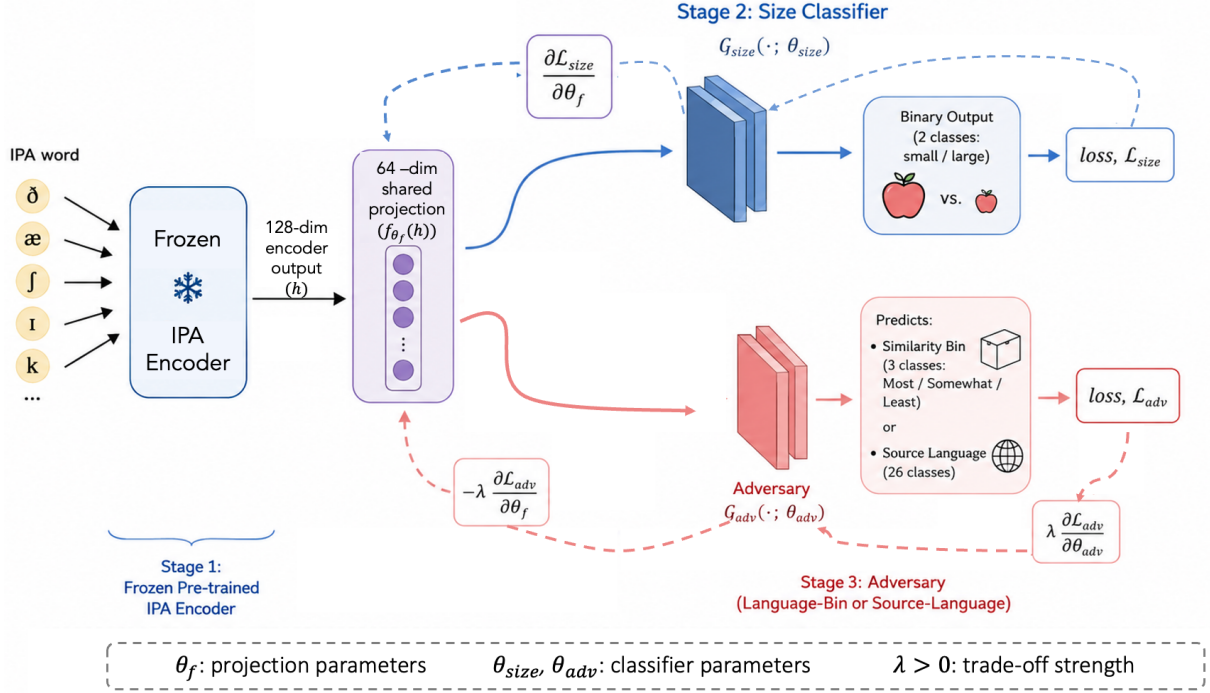


Figure 1: Adversarial training pipeline. A BERT encoder is pretrained on 1.6M IPA-transcribed words from WikiPron, then frozen as a fixed feature extractor. The embedding layer projects outputs to features that feed both a size classifier and a language-bin adversary. During backpropagation, the gradient reversal layer inverts gradients from the adversary, so that the projection learns features that support size prediction while making similarity-bin/source-language cues less usable by the adversarial head.

tor over IPA segment counts. Because standard tokenizers mishandle diacritics and other IPA symbols, we use *ipatok* (pavelsof and Martinović, 2025) for phoneme segmentation. We exclude suprasegmentals such as tone, stress, and intonation because their encoding is highly language-specific and would require native phonetic expertise for consistent annotation.

We use a leave-one-language-out (LOLO) design. For each held-out target language ℓ , we rank the remaining 26 languages by LDN similarity to ℓ and divide them into three tertiles: *Most Similar*, *Somewhat Similar*, and *Least Similar*. We train separate Logistic Regression and Decision Tree classifiers on each tertile and test on ℓ , repeating this procedure for all 27 target languages and reporting bin-wise means with 95% confidence intervals. As a stricter check on the Indo-European skew in the sample, we also train on all non-Indo-European languages and test on Indo-European languages, and vice versa.

Decision Trees use maximum depth 6 and minimum-samples-split values between 2 and 10 to reduce overfitting. Logistic Regression uses the lbfgs solver with L2 regularization ($C = 1.0$) and

a maximum of 1,000 iterations. Our primary evaluation is held-out accuracy relative to chance. We assess significance against the 50% chance baseline. For interpretability, we analyze Logistic Regression coefficients to identify phonemes associated with smallness or largeness, and Decision Tree root splits to identify the highest-gain phoneme feature for each target language.

3.3 Adversarial Scrubber

The adversarial models test whether size-predictive information remains usable when similarity-bin or source-language cues are made harder for the trained model to exploit (Figure 1). Because our curated dataset is small, we first pretrain a Transformer encoder on 1.6M IPA-transcribed WikiPron entries (Lee et al., 2020) using masked language modeling. This gives the model broad phonotactic knowledge before downstream training. WikiPron covers 26/27 languages in our sample, with entry counts ranging from 513 to 466,668. Five languages have fewer than 10K entries, 12 have more than 50K, and Twi Asante is absent (Appendix D).

The encoder has 2 layers, 128 hidden units, and a 115-character IPA vocabulary; after 2 epochs, it is frozen and used as a fixed feature extractor. On top

of the frozen encoder, we learn a 64-dimensional projection and two task heads: a binary size classifier and a multi-class adversary. For the similarity-bin adversary, each training word is labeled by the similarity tertile of its source language relative to the held-out target language. The architecture is:

- **Encoder/projection (E):** a linear layer (128→64) with dropout, receiving gradients from both task heads.
- **Size classifier (C):** a two-layer feed-forward network with ReLU and dropout, followed by a softmax over *small* and *large*.
- **Adversary (A):** a two-layer feed-forward network with ReLU and dropout, followed by a softmax over similarity bins, connected to E through a gradient-reversal layer.

The model is trained with the minimax objective:

$$\min_{E,C} \max_A \mathcal{L}_{\text{size}}(C(E(x)), y_{\text{size}}) - \lambda \cdot \mathcal{L}_{\text{bin}}(A(E(x)), y_{\text{bin}}) \quad (1)$$

Here, $\mathcal{L}_{\text{size}}$ and $\mathcal{L}_{\text{bin}}$ are cross-entropy losses. The gradient-reversal layer multiplies adversary gradients by $-\lambda$ before they reach E , so the projection is trained to preserve size-predictive information while reducing the jointly trained adversary’s access to bin information. We ramp λ from 0 to 1 over 20 epochs following Ganin and Lempitsky (2015), apply dropout ($p = 0.1$), and compare against a $\lambda = 0$ baseline. Chance accuracy is 50% for size classification and 33.3% for the 3-way bin adversary.

Our main adversarial evaluation reports two quantities: held-out size accuracy, which should remain above chance if size cues survive, and training-bin accuracy, which should approach chance if the jointly trained adversary cannot recover similarity-bin labels. Because bin labels may miss exact source-language cues, we also run a stricter 26-way source-language adversary with the same LOLO setup and training schedule, replacing the bin head with a head that predicts the training language. Finally, we train a post-hoc linear probe on the learned embeddings to test whether source-language identity remains recoverable even when it is difficult for the jointly trained model to use.

4 Results

4.1 Baseline Classification Accuracy

Baseline classifiers detect size information above chance, though the signal weakens with linguistic distance. Decision Trees performed significantly above chance in all three bins (Most: 60.2%, Somewhat: 56.4%, Least: 55.8%; $p < 0.01$), with medium-to-large effect sizes (Cohen’s $d = 1.23, 0.78, 0.65$). Logistic Regression was above chance for the Most and Somewhat Similar bins (59.3%, 57.0%; $p < 0.01$; $d = 0.73, 0.60$), but not the Least Similar bin (52.2%).

In aggregate, accuracy declines from the Most Similar to Least Similar bins for both classifiers (Figure 2), but these differences are not statistically significant under a one-way ANOVA ($p = 0.08$ for Logistic Regression; $p = 0.19$ for Decision Trees), and the pattern is not monotonic for every target. For example, Hungarian Logistic Regression accuracy increases from 20% to 27% to 37% across the three bins, while remaining below chance throughout. We therefore interpret the distance trend cautiously. The key result is not that there is a confirmed monotonic relationship between genealogical distance and accuracy, but that the size signal remains detectable in more difficult cross-family transfer settings, especially for Decision Trees. Table 1 summarizes bin-level performance, while detailed language-level accuracies are shown in Figure 2.

Indo-European holdout. As a stricter check on the baseline trend, we train on the 11 non-Indo-European languages and test on the 16 Indo-European languages. Both classifiers remain above chance: Logistic Regression reaches 57.9% ($p = .010$) and Decision Trees 57.5% ($p = .001$). In the reverse direction, performance is also numerically above chance but not significant (Logistic Regression: 52.1%, $p = .317$; Decision Trees: 54.5%, $p = .203$). The positive direction in both splits is consistent with the modest cross-family trend above, although only the non-Indo-European to Indo-European transfer is statistically significant.

4.2 Phonological Features & Size Symbolism

Stable vowel patterns provide the clearest cross-family signal. For Logistic Regression, we first filtered for phonemes with consistent coefficient direction across models ($> 60\%$ same sign), then averaged coefficients across the models in which

Bin	Model	Acc.	Prec.	Rec.	F1	F1 _{small}	F1 _{large}
Most Similar	Logistic Reg.	59.3 ^{***}	60.8	59.3	58.1	58.3	57.8
	Decision Tree	60.2 ^{***}	63.2	60.2	58.2	57.1	59.3
Somewhat Similar	Logistic Reg.	57.0 ^{**}	58.0	57.0	55.9	58.2	53.7
	Decision Tree	56.4 ^{***}	55.7	56.4	50.2	53.1	47.3
Least Similar	Logistic Reg.	52.2	52.4	52.2	50.0	49.5	50.5
	Decision Tree	55.8 ^{**}	56.5	55.8	50.7	47.5	54.0

Table 1: Classification performance (%) by model and similarity bin, averaged over 27 languages. Precision, Recall, and F1 are macro-averaged across the two classes; F1_{small} and F1_{large} report per-class F1. Means are across held-out target languages; significance is tested against 50% chance across target-language accuracies. Asterisks indicate bins where accuracy is significantly above chance (^{**} $p < 0.01$, ^{***} $p < 0.001$).

each phoneme appeared. Among vowels present in all 81 language–bin models, /o/ showed the strongest association with largeness (100% positive, mean coefficient = +0.67), while /i/ showed the strongest association with smallness (99% negative, mean = −0.44). The low/open vowel /a/ also pointed consistently toward largeness, though with a smaller mean coefficient (+0.26). Because each language contains only 15 small and 15 large adjectives, we tested whether these coefficient directions were overly dependent on the particular words retained. We refit the analysis over 1,000 bootstrap resamples within each language, sampling with replacement from the existing 30-word set. Coefficient-sign consistency was 0.82 for /i/, 0.83 for /a/, and 0.83 for /o/. Thus, the most stable cross-family pattern aligns with classic size-symbolism findings: high/front vowels such as /i/ cue smallness, while lower or back/rounded vowels such as /a/ and /o/ cue largeness.

Decision Trees corroborate the vowel-dominant pattern. We next examined the root split of each Decision Tree, i.e., the feature that best separates large from small words at the top of the tree. Across 81 language–bin combinations, vowels dominated: /a/ was the root split in 22 cases (27%) and /i/ in 19 (24%). Consonants appeared less often, including fricatives (/h/: 5, /v/: 4, /f/: 3, /s/: 3), nasals (/m/: 4), and stops (/k/: 3). This pattern held across similarity bins: /i/ and /a/ accounted for 63% of root splits in the Most Similar bin and 55% in the Least Similar bin.

Models with /a/ as the root split exceeded chance in 16/22 cases, with mean accuracy significantly above chance ($M = 0.57$, $t(21) = 3.38$, $p < 0.01$). Models with /i/ exceeded chance in

15/19 cases ($M = 0.58$, $t(18) = 4.52$, $p < 0.001$). These root splits also span diverse families: /a/ appears as the top split for Turkish, Mandarin, Hebrew, Italian, and Polish, while /i/ appears for Marathi, Spanish, Greek, and Hungarian.

Some high-coefficient phones are informative but more language-conditioned. The largest Logistic Regression coefficients often came from phones that appeared only in a subset of languages. For largeness, /fi/ had a mean coefficient of +1.32 across 37 models, and /ẽ/ had +1.12 across 26 models. For smallness, /ã/ had a mean coefficient of −0.86 across 26 models, and /ñ/ had −0.85 across 35 models. Because these phones are not present across the full sample, we interpret them as language- or family-conditioned cues rather than broad cross-family evidence.

Overall, both classifiers agree that /i/ is a central cue for smallness. For largeness, Decision Trees emphasize the low/open vowel /a/, while Logistic Regression assigns stronger weight to the back/rounded vowel /o/. The clearest cross-family signal is therefore vowel-dominant, with consonants providing additional but less stable cues.

4.3 Adversarial Training Results

We trained two model variants for each of the 27 target languages: (1) adversarial (λ ramped from 0 to 1), and (2) baseline ($\lambda = 0$, no adversarial pressure). For each variant, we conducted 5 runs (Figure 3).

Epoch selection. For the adversarial condition, we selected the epoch (greater than or equal to 8, when $\lambda \approx 1$) that maximized the difference between train size accuracy and train bin accuracy,

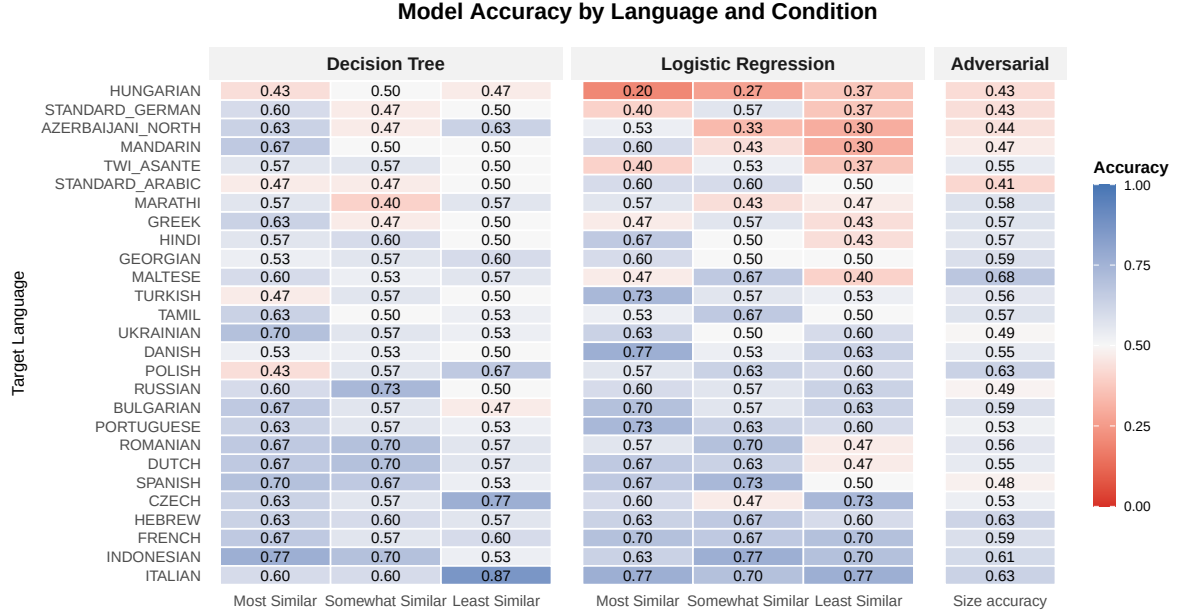


Figure 2: Baseline performance weakens with increasing dissimilarity, with several languages dropping to chance in the Least Similar bin. Under the similarity-bin adversary, above-chance size prediction is maintained in 15/25 languages with non-zero variance, suggesting a small but consistent cross-family size signal under genealogical pressure.

with train bin accuracy at or above chance (33.3%; if no such epoch existed, we removed this constraint). This identifies the point where bin suppression is strongest while size information is preserved. For the baseline ($\lambda = 0$), we selected the epoch (greater than 10) with the highest train bin accuracy.

The bin adversary did not predict similarity bins above chance: train bin accuracy was 34.0% (vs. 54.4% baseline; 33.3% chance), while held-out size accuracy was 54.4% (vs. 55.4% baseline). The 4.4-point gain over chance is an 8.8% relative error reduction over a 50% baseline. Our aim is not to maximize accuracy, but to test whether size information persists once family cues are suppressed. A paired t -test across languages showed no significant difference in size prediction between language-suppressed and non-suppressed conditions ($t(26) = 1.76, p = .09$).

For each language, we averaged over 5 runs and used one-sample t -tests. Size accuracy was tested against 50% chance with a one-tailed test and Benjamini-Hochberg FDR correction across the 25 languages with non-zero variance. Mandarin and Ukrainian had identical accuracy across all five runs, so the one-sample t -test was undefined for these two languages. Bin accuracy was tested against 33.3% chance with a two-tailed test. We

treat non-significantly-above-chance bin accuracy as evidence that the jointly trained adversary cannot reliably recover bin labels, which we report at $p \geq 0.01$. Of the remaining 25, 15 (60%) showed above-chance size accuracy after BH correction ($q < 0.05$) while bin accuracy was not significantly above chance: Bulgarian, Dutch, French, Georgian, Greek, Hebrew, Hindi, Indonesian, Italian, Maltese, Marathi, Polish, Portuguese, Romanian, and Turkish. These span 9 of the 13 language families in our dataset.

Direct source-language suppression preserves size prediction. Under the stricter 26-way source-language adversary, training-language accuracy dropped from 23.9% to 5.1%, near chance (3.8%), while held-out size accuracy was unchanged (55.8% adversarial vs. 54.2% baseline; difference not significant). A post-hoc linear probe recovered source-language identity from the learned embeddings at 43.4% accuracy, indicating that language information remained. This aligns with prior work on adversarial representation learning: adversaries can hinder the jointly trained model from using protected information without fully removing it for separate probes (Elazar and Goldberg, 2018; Ravfogel et al., 2020). We therefore do not claim fully language-invariant representations, but show that size prediction is stable even when exact language

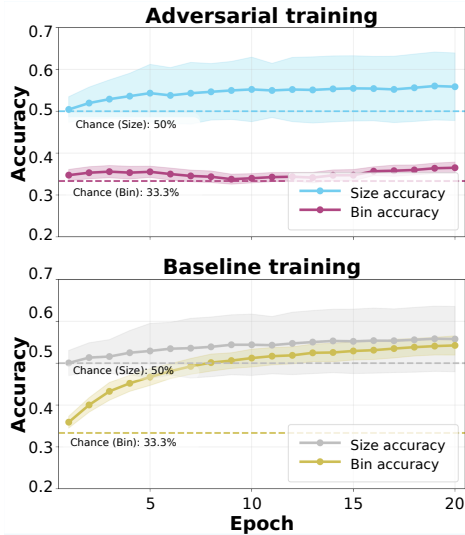


Figure 3: Mean accuracy across 27 languages over 20 epochs. Shaded regions indicate ± 1 standard deviation across languages. Without adversarial pressure, bin accuracy steadily increases, indicating the model learns language-specific cues. Under adversarial training, bin accuracy approaches chance for the jointly trained adversary, while size accuracy remains above chance and similar to baseline, suggesting the adversary suppresses language cues without affecting size prediction.

ID is explicitly suppressed during training.

4.4 Ablation Experiments

We show several ablation experiments in Table 2. All ablations are run on the interpretable bag-of-phoneme classifiers rather than the neural adversarial model, since our goal is to isolate which phonological cues support cross-family transfer.

Scrambled size labels. To test whether models capture phoneme–meaning associations rather than dataset artifacts, we randomly shuffled test-set size labels within each language while preserving class balance. Both classifiers fell to chance across all bins: Logistic Regression achieved 48.6%, 47.9%, and 50.0% for the Most, Somewhat, and Least Similar bins, while Decision Trees achieved 50.9%, 50.0%, and 50.4%, respectively. None were significant. This confirms that above-chance performance depends on the correct size labels.

No vowels. Since vowels were the strongest cues in the baseline models, we removed them from the features and retrained the classifier models. Logistic Regression stayed above chance only in the Most Similar bin (58.6%), dropping to 52.8% and 51.1% in the Somewhat and Least Similar bins; Decision Trees showed a similar pattern (55.8%,

54.6%, 52.2%). Thus, consonants retain some signal in near-family settings, but vowels carry the most reliable cross-family cues.

High-frequency phones only. For this analysis, we limited the models to 23 cross-linguistically common phonemes chosen with our collaborating phonetician. We observed that both classifiers performed close to the full model. This suggests that much of the size signal can be retained using typologically frequent segments rather than rare, language-specific phones, possibly aided by a regularization effect from removing sparse features.

Plosives vs. nasals. For this analysis, we compared classifier performance from models trained on plosives and nasals, two frequent consonant classes with distinct articulatory profiles: plosives involve oral closure with an audible release, while nasals involve continuous nasal airflow. Plosives outperformed nasals in both models (54.5% vs. 51.6% for Logistic Regression; 56.4% vs. 52.5% for Decision Trees).

Romance morphology control. Because bag-of-phoneme features count all segments, they may conflate lexical stems with grammatical endings. This matters for Spanish and Portuguese, where masculine singular adjectives often end in /o/, potentially inflating the apparent /o/–largeness association. To test this, we reran the logistic-regression coefficient analysis excluding Spanish and Portuguese. The pattern remained: /o/ was positive in all remaining models (mean coefficient = 0.48; 75/75 models), while /i/ remained strongly associated with smallness (mean coefficient = -0.43; negative in 98.7% of models). Therefore, this suggests that the vowel–size pattern is not driven solely by Spanish/Portuguese inflectional morphology.

5 Discussion

Interpreting phonetic trends. Prior work links back or low vowels to largeness and front vowels like /i/ to smallness via acoustic cues: lower-frequency vowels suggest larger resonating cavities, while higher-frequency vowels suggest smaller ones (Sapir, 1929; Newman, 1933; Ohala, 1994; Shinohara and Kawahara, 2010). Our results are consistent with this pattern: across models, /o/ best predicts largeness, /i/ smallness, and low/open /a/ also trends toward largeness. These widely attested vowels yield the clearest cross-family signal. Ablations support a vowel-dominant interpre-

ID	Ablation Condition	LR Mean	DT Mean	LR DT by Bin
0	Baseline (all phonemes)	56.1	57.5	59.3 / 57.0 / 52.2 60.2 / 56.4 / 55.8
1	Scrambled Labels	48.8	50.4	48.6 / 47.9 / 50.0 50.9 / 50.0 / 50.4
2	No Vowels	54.2	54.2	58.6 / 52.8 / 51.1 55.8 / 54.6 / 52.2
3	High-Frequency Phonemes	56.5	58.5	58.5 / 55.6 / 55.4 60.2 / 57.8 / 57.4
4	Only Plosives	54.5	56.4	56.7 / 53.5 / 53.2 58.6 / 55.9 / 54.6
5	Only Nasals	51.6	52.5	52.0 / 51.0 / 51.8 53.0 / 51.6 / 53.0

Table 2: Mean bin-wise accuracy (%) for logistic regression (LR) and decision tree (DT) classifiers across ablation conditions. Bin results are shown in the format: Most / Somewhat / Least Similar. Full phoneme sets in [Appendix G](#).

tation. Removing vowels preserves some signal in the closest similarity bin but weakens transfer to more distant languages, suggesting that consonants offer less stable, more local cues. High-coefficient consonants such as /f/ are informative in subsets of languages, but because they are not consistently present across the sample, we treat them as language- or family-specific rather than broad cross-family evidence. Because our analysis uses symbolic phonological forms rather than acoustic measurements or perceptual judgments, this alignment with prior frequency-code accounts does not establish the underlying acoustic or perceptual mechanism.

Interpreting the adversarial results. The adversarial experiments test whether size information remains usable when models are discouraged from using genealogy or source-language shortcuts. In both settings, the pattern is similar: the bin adversary could not reliably predict similarity bins, and the stricter source-language adversary pushed source-language prediction near chance, while held-out size prediction stayed above chance. Languages meeting both criteria (above-chance size accuracy and non-significant similarity-bin accuracy) spanned 9 of the 13 families. These results do not imply full removal of language information. A post-hoc linear probe still recovers source-language identity, matching prior work showing that adversarial training can reduce the *usability* of protected information without guaranteeing its *removal* (Elazar and Goldberg, 2018; Ravfogel et al., 2020). We therefore treat the adversary as a stress test showing that size information remains detectable even when language identity is difficult for the model to exploit.

Across the distance-stratified classifiers, phoneme analyses, ablations, and both adversarial settings, we find the same modest but consistent residual signal. Additionally, the Indo-European

holdout shows significant transfer from non-Indo-European to Indo-European languages, while the reverse direction remains above chance but is not significant. This suggests that the signal is unlikely to be explained solely by the Indo-European skew in our sample, while its asymmetry indicates that it is neither uniform nor large. Stronger claims require broader language coverage, multiple varieties per language, richer phonetic modeling, and human perceptual validation.

6 Conclusion

We frame size sound symbolism as a stress test: what survives when genealogical and source-language shortcuts are made difficult to use? Across 27 languages and 13 families, we find a small but detectable cross-family size signal, with the clearest vowel cues consistent with established accounts of size symbolism. Consonants add weaker, more language-conditioned evidence. Adversarial experiments show that size prediction remains above chance when similarity-bin and source-language cues are discouraged, though post-hoc probes confirm that language identity is not fully removed. Thus, we do not claim a universal sound-meaning mapping; rather, traces of size symbolism persist under genealogy-aware cross-linguistic controls. More broadly, our framework combines distance-aware transfer, interpretable phonological features, ablations, adversarial stress tests, and residual-information probes to strengthen evidence for cross-linguistic size symbolism.

Limitations

A key limitation of this study is the modest size of our dataset, which reflects the practical difficulty of conducting cross-linguistic research at scale. Recruiting native speakers for 27 languages, particularly those languages with limited academic or online presence, posed a significant logistical chal-

lenge. Ideally, all transcriptions would be vetted by native speakers with formal linguistic training, but such individuals are extremely rare for many of the languages in our sample. Measurement noise may attenuate true effects but can also introduce language-specific artifacts, so differences need to be interpreted cautiously. Our empirical evaluation focuses on size symbolism, where we find a modest but detectable cross-family signal. Whether the same framework generalizes to other iconic domains, such as shape, texture, or affect, remains untested. While we aimed for genealogical diversity using WALS and included languages from 13 families, the dataset remains slightly skewed toward Indo-European languages due to the relative accessibility of speakers. Additionally, we relied on phonemic rather than phonetic transcriptions; although phonemic representations offer greater cross-speaker consistency, they may miss subtle sound-symbolic cues present in actual pronunciation. Bag-of-phoneme features also discard segment position and do not distinguish stems from affixes; the Romance control addresses one specific suffixal explanation, but not this broader limitation. Suprasegmentals are only partially incorporated in our analysis. Full prosodic modeling would benefit some languages (e.g., tone systems), but suprasegmental features vary across languages and require native-speaker phonetic expertise that is difficult to obtain, particularly for lower-resourced languages. Finally, larger samples covering a broader range of languages, as well as perceptual experiments, may be needed to establish how broadly these size-symbolic patterns generalize.

Acknowledgments

We thank Christopher Lucas (SOAS, University of London), Michael Flier (Harvard University), Katya Pertsova (UNC Chapel Hill), Julie De Lataulade de Laas (UNC Chapel Hill), Paul Roberge (UNC Chapel Hill) and Nina Topintzi (Aristotle University of Thessaloniki) for their expertise in providing feedback on our word lists and IPA transcriptions for Maltese, Ukrainian, Russian, French, German, and Greek.

Generative AI Statement

We utilized AI tools for polishing original written content, writing and debugging experiments, performing statistical analysis, and generating diagrams.

References

- Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- E. Annamalai. 1968. Onomatopoeic resistance to sound change in dravidian. In *Studies in Indian Linguistics: Professor M. B. Emeneau Sasitipūrti Volume*, pages 15–19. Deccan College and Annamalai University, Poona.
- Dik Bakker, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. [Adding typology to lexico-statistics: A combined approach to language classification](#). *Linguistic Typology*, 13(1):169–181.
- Brent Berlin. 1994. Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. *Sound symbolism*, pages 76–93.
- Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. [Sound–meaning association biases evidenced across thousands of languages](#). *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.
- Dwight L Bolinger. 1940. Word affinities. *American speech*, 15(1):62–73.
- Roger W Brown, Abraham H Black, and Arnold E Horowitz. 1955. Phonetic symbolism in natural languages. *The Journal of Abnormal and Social Psychology*, 50(3):388.
- Ricardo Calvillo-Torres, Jorge Haro, Pilar Ferré, Claudia Poch, and José A. Hinojosa. 2024. [Sound symbolic associations in spanish emotional words: affective dimensions and discrete emotions](#). *Cognition and Emotion*, pages 1–17. Epub ahead of print.
- G. Tucker Childs. 1994. African ideophones. In Leanne Hinton, Johanna Nichols, and John J. Ohala, editors, *Sound Symbolism*, pages 178–204. Cambridge University Press, Cambridge.
- G. Tucker Childs. 2008. Research on ideophones, whither hence?: The need for a social theory of ideophones. In F. K. Erhard Voeltz and Christa Kilian-Hatz, editors, *Ideophones*, pages 63–73. John Benjamins Publishing Company, Amsterdam.
- David JA Clines and David M Stec. 2018. *The dictionary of Classical Hebrew revised*, volume 1. Sheffield Phoenix Press Sheffield.
- David Crystal. 2009. [The ugliest words; marilyn monroe’s dog; and wacko races](#). *The Guardian*, July 17, 2009. Accessed May 18, 2025.
- DeepL SE. 2025. [DeepL translator](#). Online. Computer software.

- Gérard Diffloth. 1972. Notes on expressive meaning. In *Chicago Linguistic Society*, volume 8, pages 440–47. Chicago, IL.
- Gérard Diffloth. 1976. Expressives in semai. *Oceanic linguistics special publications*, (13):249–264.
- Mark Dingemanse, Will Schuerman, Eva Reinisch, Sylvia Tufvesson, and Holger Mitterer. 2016. [What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages](#). *Language*, 92(2):e117–e133.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online \(v2020.4\)](#). Zenodo.
- Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.
- M. B. Emeneau. 1995. [The palatalizing rule in tamil-malayalam and telugu](#). *Journal of the American Oriental Society*, 115(3):401–409.
- Yaroslav Ganin and Victor Lempitsky. 2015. [Unsupervised domain adaptation by backpropagation](#). In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.
- Google. 2025. [Google translate](#). Online. Computer software.
- E Dario Gutiérrez, Roger Levy, and Benjamin Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2379–2388.
- Shoko Saito Hamano. 1986. *The sound-symbolic system of Japanese (ideophones, onomatopoeia, expressives, iconicity)*. University of Florida.
- Otto Jespersen. 1922. Symbolic value of the vowel i. *Philologica*, 1.
- Niklas Erben Johansson and Jordan Zlatev. 2013. [Motivations for sound symbolism in spatial deixis: A typological study of 101 languages](#). *Public Journal of Semiotics*, 5(1):3–20.
- Alexander James Kilpatrick, Aleksandra Ćwiek, and Shigeto Kawahara. 2023. [Random forests, sound symbolism and pokémon evolution](#). *PLOS ONE*, 18(1):e0279350.
- Wolfgang Köhler. 1929. *Gestalt Psychology*. Liveright, New York.
- Paul A Kotey. 1998. *Twí-English/English-Twí Dictionary*. Hippocrene Books.
- Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. [Massively multilingual pronunciation modeling with WikiPron](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.
- Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zita Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint arXiv:2009.13028*.
- Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. [With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. [PanPhon: A resource for mapping IPA segments to articulatory feature vectors](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.
- Stanley S Newman. 1933. Further experiments in phonetic symbolism. *The American Journal of Psychology*, 45(1):53–75.
- Johanna Nichols. 1971. Diminutive consonant symbolism in western north america. *Language*, pages 826–848.
- Svend V. Nielsen. 2017. [Mds plot of the world’s languages](#). Accessed: 2025-07-25.
- John J. Ohala. 1994. The frequency code underlies the sound-symbolic use of voice pitch. In Leanne Hinton, Johanna Nichols, and John J. Ohala, editors, *Sound Symbolism*, pages 325–347. Cambridge University Press, Cambridge.
- Ananya Passi and S. P. Arun. 2024. [The bouba–kiki effect is predicted by sound properties but not speech properties](#). *Attention, Perception, & Psychophysics*, 86(3):976–990.
- pavelsof and Viktor Martinović. 2025. [pavelsof/ipatok](#). <https://github.com/pavelsof/ipatok>.
- Fabio Petroni and Maurizio Serva. 2010. [Measures of lexical distance between languages](#). *Physica A: Statistical Mechanics and its Applications*, 389(23):2280–2283.
- Tiago Pimentel, Arya D McCarthy, Damián E Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. *arXiv preprint arXiv:1906.05906*.

- Tiago Pimentel, Brian Roark, Søren Wichmann, Ryan Cotterell, and Damián Blasi. 2021. Finding concept-specific biases in form–meaning associations. *arXiv preprint arXiv:2104.06325*.
- Uriel Cohen Priva, Emily Strand, Shiyang Yang, William Mizgerd, Abigail Creighton, Justin Bai, Rebecca Mathew, Allison Shao, Jordan Schuster, and Daniela Wiepert. 2021. The cross-linguistic phonological frequencies (xpf) corpus. <https://github.com/CohenPr-XPF/XPF>.
- A. Raevskiy, S. Sakamoto, and N. Sakai. 2023. *Psychoacoustic study of japanese mimetics for food textures*. In K. Kondo, M.F. Horng, J.S. Pan, and P. Hu, editors, *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, volume 339 of *Smart Innovation, Systems and Technologies*. Springer, Singapore.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics.
- Richard Rhodes and John M Lawler. 1981. Athematic metaphors. In *Chicago Linguistic Society*, volume 17, pages 318–342.
- William J. Samarin. 1965. Perspective on african ideophones. *African Studies*, 24(2):117–121.
- Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12:225–239.
- Ferdinand de Saussure. 2011. *Course in General Linguistics*. Columbia University Press, New York. Original work published 1916.
- Kazuko Shinohara and Shigeto Kawahara. 2010. A cross-linguistic study of sound symbolism: The images of size. In *Annual meeting of the berkeley linguistics society*, pages 396–410.
- David M. Sidhu and Penny M. Pexman. 2019. *The sound symbolism of names*. *Current Directions in Psychological Science*, 28(4):398–402.
- David M. Sidhu, Chris Westbury, Geoff Hollis, and Penny M. Pexman. 2021. *Sound symbolism shapes the english language: The maluma/takete effect in english nouns*. *Psychonomic Bulletin & Review*, 28(4):1390–1398.
- Tripod Technology GmbH. 2025. *Tandem: Language exchange*. App Store. Mobile application software.
- Russell Ultan. 1978. Size-sound symbolism. *Universals of human language*, 2:525.
- Hans Wehr and J Milton Cowan. 1994. Arabic-english dictionary: The hans wehr dictionary of modern written arabic. *Urbana: Spoken Language Services*, 1(2):3.
- Bodo Winter and Marcus Perlman. 2021. *Size sound symbolism in the english lexicon*. *Glossa: a journal of general linguistics*, 6(1):79.
- Bodo Winter, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. 2022. Trilled/r/is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1):1035.
- Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sokratis Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, and Bodo Winter. 2022. *The bouba/kiki effect is robust across cultures and writing systems*. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390.

A Languages in the Dataset and their Levenshtein distances

Figure 4 visualizes the pairwise normalized Levenshtein distances (LDN) between the 27 languages in our dataset as a hierarchically clustered heatmap, with colored bars indicating the genealogical groupings used in our analysis. LDN quantifies lexical similarity from 40-word Swadesh lists; smaller values denote greater similarity, allowing each held-out target language to be compared against training languages at graded genealogical distances.

B English Source Adjective List

Table 3 shows the English source adjective list used to guide translation into the target languages. Because not all English adjectives have a direct or natural translation in every target language, the source list is intentionally larger than the final 30-item per-language sets.

Small-seeming adjectives	Large-seeming adjectives
Atomic, Bit, Brief, Compact, Crumb, Delicate, Diminutive, Dot, Dwarf, Dwarfish, Exiguous, Few, Fine, Inadequate, Insignificant, Less, Light, Lilliputian, Limited, Little, Meager, Meticulous, Microscopic, Mini, Miniature, Minimal, Minuscule, Minor, Minute, Modest, Narrow, Negligible, Petite, Petty, Pocket-sized, Puny, Pygmy, Reduced, Rudimentary, Scarce, Short, Slender, Slim, Small, Small-scale, Sparse, Subtle, Teensy, Teeny, Thin, Tiny, Trivial, Weak, Insufficient, Undersized	Abundant, Ample, Big, Boundless, Broad, Bulky, Colossal, Colossus, Considerable, Deep, Dense, Enormous, Expansive, Extensive, Giant, Gigantic, Great, Gross, Heavy, Hefty, High, Huge, Humongous, Immense, Imposing, Impressive, Kingly, Large, Lofty, Long, Macroscopic, Majestic, Massive, Mighty, Monstrous, Monumental, Oversized, Overwhelming, Plenty, Princely, Spacious, Sprawling, Substantial, Tall, Thick, Towering, Tremendous, Vast, Voluminous, Wide, Widespread

Table 3: English source adjective list used for translating size-related adjectives into target languages. The list is intentionally larger than the final per-language set because some English adjectives do not have direct or natural translations in every target language.

C LDN vs LDND: Choice of Genealogical Distance Measure

C.1 Background and Definitions

We use the Levenshtein Distance Normalized (LDN) measure of Petroni and Serva (2010) to quantify lexical distance between languages. For two words α_i and β_j representing meanings i and

j in languages α and β , the normalized word-level distance is defined as

$$D(\alpha_i, \beta_j) = \frac{D_l(\alpha_i, \beta_j)}{L(\alpha_i, \beta_j)}, \quad (2)$$

where D_l is the standard Levenshtein edit distance and L is the character length of the longer of the two words. The language-level LDN is then the average over all same-meaning word pairs in a Swadesh list of size M :

$$D(\alpha, \beta) = \frac{1}{M} \sum_i D(\alpha_i, \beta_i). \quad (3)$$

Bakker et al. (2009) proposed a further-normalized variant, LDND, intended to correct for accidental phonological similarity between unrelated languages. It divides LDN by a global distance term Γ that averages word distances across *different-meaning* pairs in the two languages:

$$\Gamma(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j} D(\alpha_i, \beta_j), \quad (4)$$

$$D_s(\alpha, \beta) = \frac{D(\alpha, \beta)}{\Gamma(\alpha, \beta)}. \quad (5)$$

The motivation is that Γ captures background similarity in character distributions independent of meaning. Dividing by it, in theory, should remove similarity that might be “noise” (due to shared phonological inventory rather than shared ancestry).

C.2 Why LDN over LDND

The LDND correction is built on the assumption that Γ captures only accidental similarity — for example, two unrelated languages that happen to share sound inventories will have small Γ for reasons unrelated to ancestry. If this assumption holds, Γ should be uncorrelated with lexical distance between languages, and dividing it out cleanly removes a nuisance term.

Petroni and Serva (2010) test this directly by computing the correlation between Γ and D_s across all language pairs in a sample. They find $C(\Gamma, D_s) = 0.55$ for Indo-European and 0.40 for Austronesian (both positive). In other words, pairs that look closely related under LDND also tend to have small Γ . This means Γ is not just noise. Dividing by it therefore removes real information about language relatedness rather than isolating accidental similarity, which is why Petroni and Serva (2010) recommend LDN over LDND.

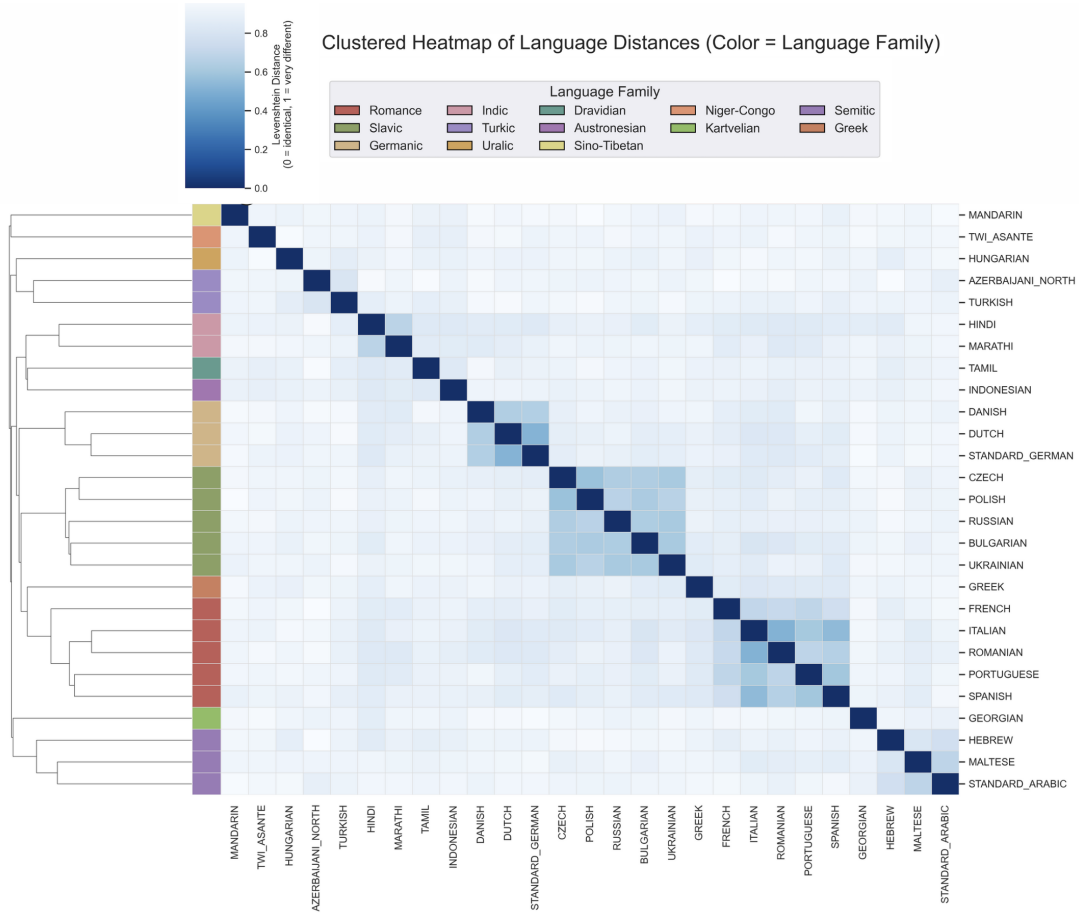


Figure 4: Levenshtein distance heatmap showing pairwise lexical similarity between the 27 languages in our dataset. Languages are grouped hierarchically and colored bars indicate language families.

C.3 Replication on Our 27-Language Sample

We repeated this analysis on our dataset. Using the same ASJP-derived Swadesh data underlying our LDN tertiles, we computed D , Γ , and D_s for all 351 unique language pairs and found $C(\Gamma, D_s) = 0.49$. Petroni and Serva (2010) computed their values within single language families, whereas ours are computed across pairs from 13 families pooled together. The correlation is comparable in magnitude to theirs despite the greater diversity, suggesting Γ continues to carry information about language relatedness in our sample. Using LDND would attenuate this signal, which is why we choose LDN over LDND.

C.4 Agreement Between LDN and LDND Rankings

To verify that the choice of metric is not driving our results, we computed the per-language Spearman rank correlation between LDN and LDND distances. For each target language, we ranked the other 26 languages by LDN and by LDND and

computed the correlation between the two rankings. Results are shown in Table 4.

Language	ρ	Language	ρ
Azerbaijani North	0.817	Maltese	0.819
Bulgarian	0.961	Mandarin	0.764
Czech	0.963	Marathi	0.914
Danish	0.990	Polish	0.960
Dutch	0.985	Portuguese	0.939
French	0.968	Romanian	0.977
Georgian	0.772	Russian	0.959
Greek	0.809	Spanish	0.948
Hebrew	0.640	Standard Arabic	0.594
Hindi	0.825	Standard German	0.963
Hungarian	0.636	Tamil	0.566
Indonesian	0.605	Turkish	0.714
Italian	0.971	Twi Asante	0.621
		Ukrainian	0.934

Table 4: Per-language Spearman rank correlation between LDN and LDND distances over the other 26 languages.

Across all 27 languages, the mean correlation is $\rho = 0.84$ (SD = 0.14), with a minimum of 0.57 (Tamil) and a maximum of 0.99 (Danish). Com-

binmed with the $C(\Gamma, D_s)$ result above, this supports our use of LDN as the metrics rank language pairs similarly.

D WikiPron Coverage

The phoneme encoder was pretrained on an earlier WikiPron snapshot containing approximately 1.6M IPA transcriptions. Because that archived TSV does not retain sufficient language metadata to reconstruct exact per-language counts, we use a later pinned WikiPron repository snapshot (commit 6de2262) to approximate relative coverage (Table 5). These counts indicate relative representation rather than exact pretraining exposure.

Coverage	N	Coverage	N
< 5K	3	10K–50K	9
5K–10K	2	> 50K	12
Absent	1		

Table 5: Approximate WikiPron coverage across the 27 languages in our dataset. Five languages have fewer than 10,000 entries, 12 have more than 50,000, and Twi Asante is absent.

The < 5K group contains Azerbaijani North, Danish, and Marathi; the 5K–10K group contains Hebrew and Romanian. Azerbaijani North is approximated using the available Azerbaijani file. For Portuguese and Spanish, the multiple regional WikiPron files are used only to determine coverage-range assignment.

E Baseline training plots by language

This appendix presents per-language training dynamics for the baseline condition ($\lambda = 0$), in which no adversarial pressure is applied. Figure 5 plots both size-prediction accuracy and similarity-bin accuracy across 20 epochs for each of the 27 target languages across 5 runs. Without adversarial scrubbing, bin accuracy rises steadily above chance, indicating that the model increasingly exploits language-specific cues during training. These plots provide the per-language baseline against which the adversarial dynamics in Appendix F should be compared.

F Detailed adversarial training plots by language

This appendix shows the corresponding training dynamics under adversarial scrubbing, where λ is

ramped from 0 to 1 over the first 20 epochs. Figure 6 plots adversarial size-prediction accuracy and similarity-bin accuracy for each of the 27 target languages across 5 runs. Compared with the baseline curves in Appendix E, bin accuracy here is pushed toward chance for the jointly trained adversary while size accuracy remains above chance, illustrating that language cues are suppressed without hampering size prediction. The per-language view also reveals heterogeneity, with some languages retaining a clear size signal under adversarial pressure and others falling closer to chance.

G Ablation Phone Sets

Figure 7 specifies the phoneme inventories used in each ablation experiment reported in Section 4.4 and Table 2. For every ablation condition, we list the full set of IPA segments retained as features. The high-frequency set in particular reflects the 23 cross-linguistically common phonemes selected with our collaborating phonetician.

H Token Coverage by Language

We report phoneme-token coverage for each language in Figure 8, Figure 9, Figure 10, and Figure 11, showing how the IPA segments in our dataset map onto the BERT vocabulary trained on WikiPron. For every language we list the number of unique tokens and the full set of attested IPA symbols.

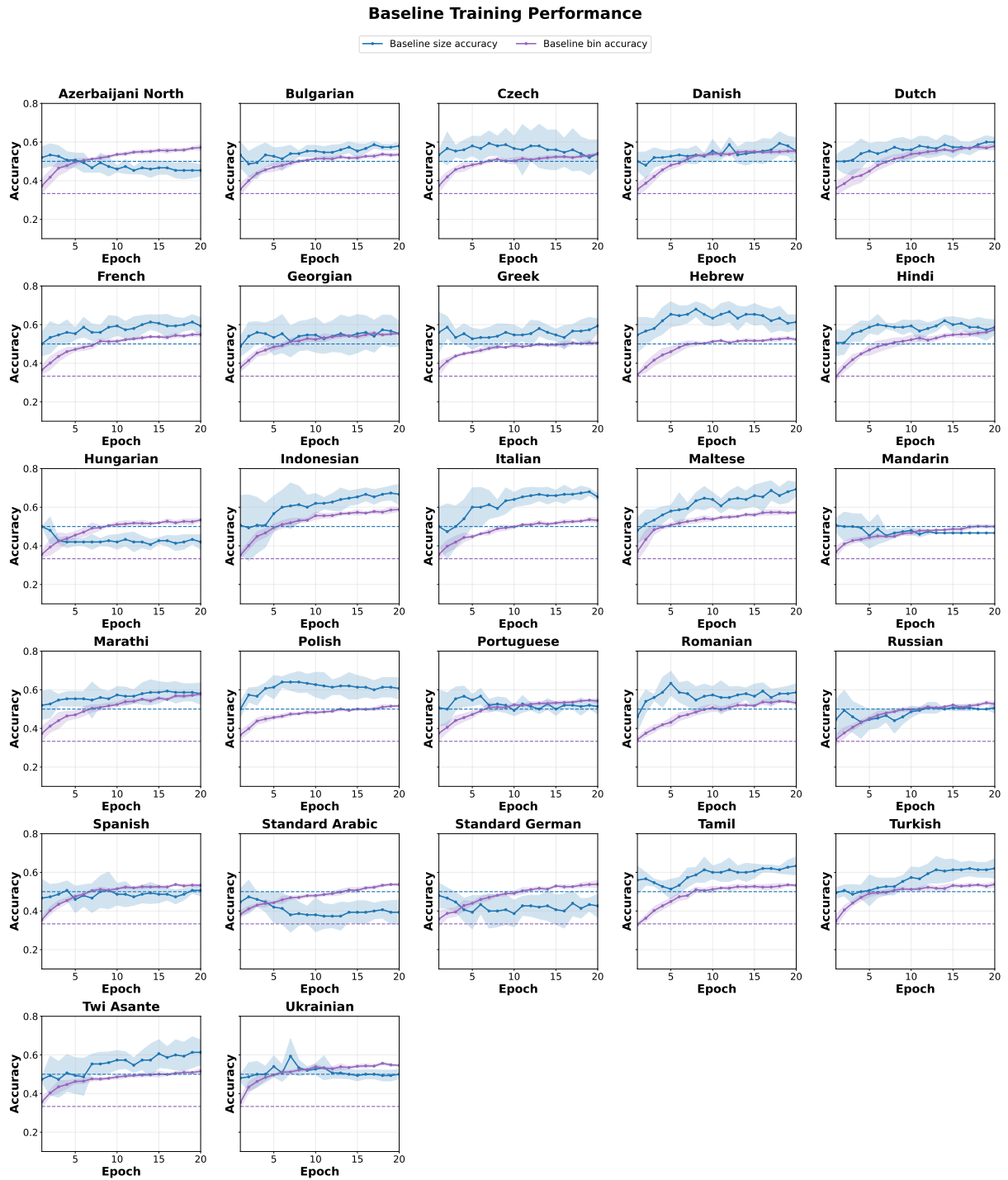


Figure 5: Baseline training dynamics ($\lambda = 0$) for each of the 27 languages. Without adversarial pressure, bin accuracy rises above chance, indicating the model learns language-specific cues. Shaded regions indicate ± 1 SD across 5 runs.

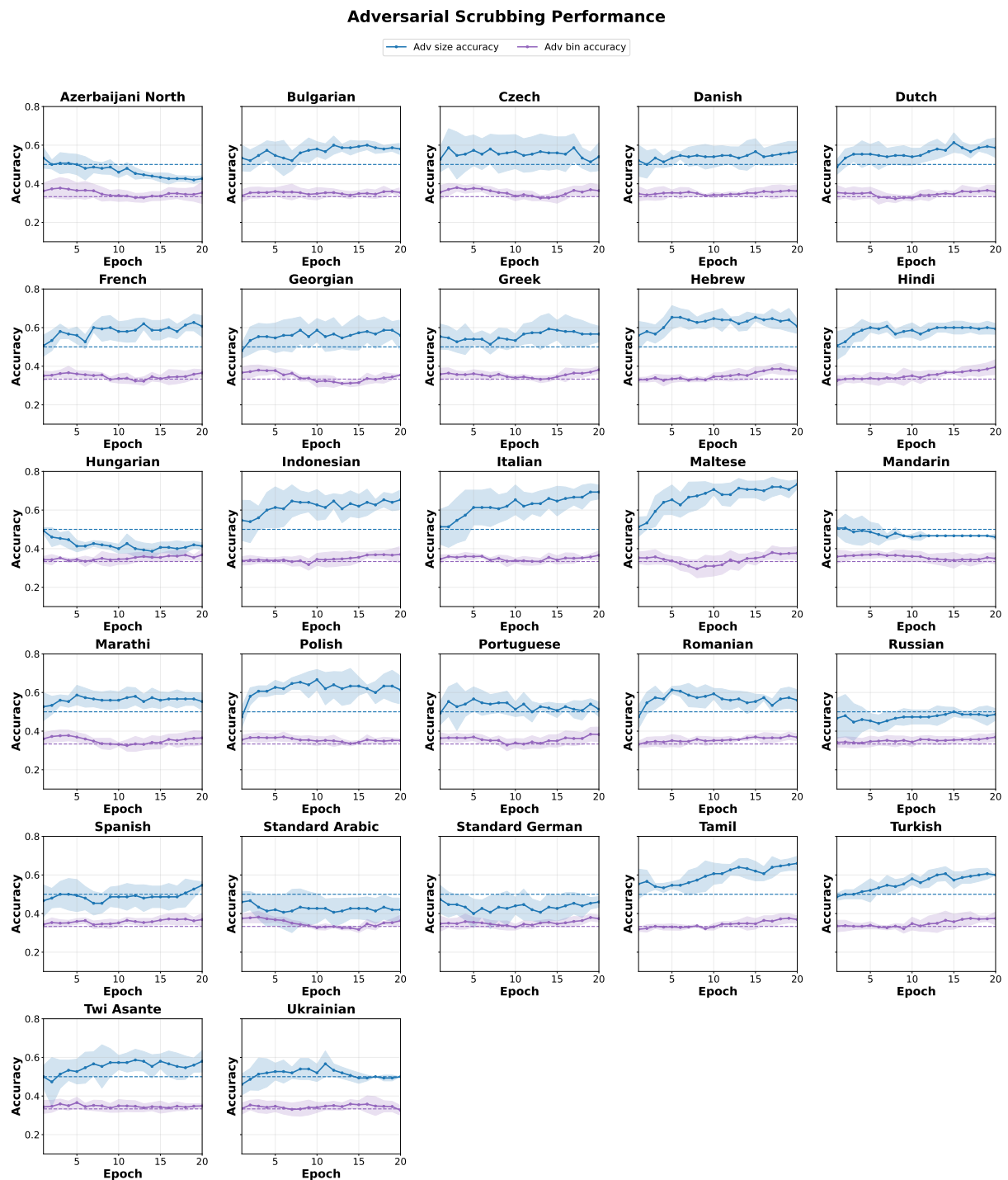


Figure 6: Adversarial training dynamics for each of the 27 languages. Shaded regions indicate ± 1 SD across 5 runs.

Language: azerbaijani_north

Unique tokens: 32

Tokens: ['b', 'd', 'd͡z', 'd͡ʒ', 'e', 'f', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 's', 't', 't͡s', 't͡ʃ', 'u', 'v', 'x', 'y', 'z', 'æ', 'œ', 'ɑ', 'ɟ', 'g', 'ɣ', 'w', 'r', 'ʃ']

Language: bulgarian

Unique tokens: 29

Tokens: ['a', 'b', 'd', 'e', 'h', 'i', 'j', 'k', 'l', 'ʎ', 'm', 'n', 'p', 'r', 's', 't', 't͡s', 't͡ʃ', 'u', 'v', 'z', 'ɐ', 'ɔ', 'ɛ', 'g', 'ɣ', 'r', 'ʃ', 'ʒ']

Language: czech

Unique tokens: 31

Tokens: ['a', 'a:', 'b', 'c', 'd', 'f', 'i', 'i:', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 't͡s', 't͡ʃ', 'u', 'u:', 'v', 'y', 'z', 'ɑ', 'ɛ', 'ɛ:', 'ɦ', 'ɲ', 'ʃ']

Language: danish

Unique tokens: 41

Tokens: ['a', 'b', 'd', 'e', 'f', 'h', 'i', 'i:', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'u', 'u:', 'v', 'y', 'y:', 'ø', 'ø:', 'œ', 'ɲ', 'ɐ', 'ɑ', 'ɑ:', 'ɔ', 'ɔ:', 'ə', 'ɛ', 'ɛ:', 'g', 'ɣ', 'ɾ', 'ʁ']

Language: dutch

Unique tokens: 39

Tokens: ['a', 'a:', 'b', 'd', 'e', 'f', 'h', 'i', 'k', 'l', 'm', 'n', 'o', 'o:', 'p', 'r', 's', 't', 't͡s', 'v', 'w', 'x', 'y', 'z', 'ø', 'ɲ', 'œ', 'ɑ', 'ɑ:', 'ɔ', 'ə', 'ɛ', 'ɛ:', 'g', 'ɣ', 'ɪ', 'ɛ', 'ʊ', 'ʏ', 'χ']

Language: french

Unique tokens: 32

Tokens: ['a', 'b', 'd', 'e', 'f', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 's', 't', 'v', 'w', 'y', 'z', 'ø', 'ø:', 'ɐ', 'ɑ', 'ɔ', 'ɔ:', 'ə', 'ɛ', 'g', 'ɣ', 'ɛ', 'ʃ', 'ʒ']

Language: georgian

Unique tokens: 29

Tokens: ['a', 'b', 'd', 'd͡z', 'h', 'i', 'j', 'k', 'l', 'l̥', 'm', 'n', 'p', 's', 't', 't͡s', 'u', 'v', 'x', 'z', 'ɑ', 'ɔ', 'ɛ', 'g', 'ɣ', 'ɦ', 'ɾ', 'r', 'ʃ']

Language: greek

Unique tokens: 25

Tokens: ['b', 'd', 'f', 'i', 'k', 'l', 'm', 'n', 'o', 'o:', 'p', 's', 't', 'u', 'v', 'x', 'ø', 'ɐ', 'ɛ', 'g', 'ɣ', 'ɲ', 'r', 'ɟ', 'θ']

Figure 8: Phoneme-token inventory of North Azerbaijani, Bulgarian, Czech, Danish, Dutch, French, Georgian, Greek.

Language: hebrew

Unique tokens: 30

Tokens: ['a', 'b', 'd', 'e', 'f', 'h', 'i', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'ts̥', 'u', 'v', 'w', 'z', 'e', 'g', 'i', 'r', 'x', 'j', 'ʔ', 'ʕ', 'χ']

Language: hindi

Unique tokens: 31

Tokens: ['b', 'bʰ', 'd', 'dʰ', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 's', 't', 'tʃ̪', 'u', 'z', 'a', 'ɔ', 'ə', 'e', 'g', 'gʱ', 'ɪ', 'ɽ', 'r', 'ʃ̪', 'ʈ', 'ʊ', 'ʋ']

Language: hungarian

Unique tokens: 41

Tokens: ['a', 'a:', 'b', 'd', 'e', 'e:', 'h', 'i', 'i:', 'j', 'k', 'l', 'm', 'n', 'o', 'o:', 'p', 'r', 's', 's:', 't', 't:', 'ts̥', 'tʃ̪', 'u', 'u:', 'v', 'y:', 'z', 'ø', 'ø:', 'ɑ', 'ɔ', 'ɛ', 'j', 'g', 'ɲ', 'ɛ̃', 'ʃ̪', 'ʒ', 'd̪']

Language: indonesian

Unique tokens: 27

Tokens: ['a', 'b', 'd', 'd͡ʒ', 'e', 'f', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'tʃ̪', 'u', 'ŋ', 'ə', 'ɛ', 'ɛ:', 'g', 'ɲ', 'r']

Language: italian

Unique tokens: 27

Tokens: ['a', 'b', 'd', 'd͡ʒ', 'e', 'f', 'h', 'i', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'tʃ̪', 'u', 'ŋ', 'ɐ', 'ə', 'g', 'ɣ', 'ɲ', 'r', 'ʎ']

Language: maltese

Unique tokens: 34

Tokens: ['a', 'b', 'b:', 'd', 'd͡ʒ', 'e', 'f', 'i', 'i:', 'j', 'k', 'l', 'm', 'm:', 'n', 'o', 'o:', 'p', 'r', 's', 's:', 't', 't:', 'ts̥', 'tʃ̪', 'u', 'u:', 'v', 'w', 'z', 'g', 'i', 'ɪ', 'ʔ']

Language: mandarin

Unique tokens: 28

Tokens: ['a', 'a:', 'e', 'i', 'k', 'm', 'n', 'p', 's', 't', 't͡ʂ', 'u', 'w', 'x', 'y', 'ŋ', 'ɑ', 'ɔ', 'ɐ', 'ə', 'ɛ', 'ɤ', 'ɪ', 'ɿ', 'ʃ̪', 'ʊ', 'z', 'ʕ']

Language: marathi

Unique tokens: 39

Tokens: ['a', 'a:', 'b', 'bʰ', 'd', 'e', 'f', 'h', 'i', 'i:', 'j', 'k', 'l', 'm', 'n', 'o', 'o:', 'p', 'r', 's', 't', 'tʃ̪', 'u', 'u:', 'z', 'ŋ', 'd̪', 'ə', 'ɛ', 'g', 'ɪ', 'ɽ', 'r', 'ʃ̪', 'ʈ', 'ʊ', 'ʋ']

Figure 9: Phoneme-token inventory of Hebrew, Hindi, Hungarian, Indonesian, Italian, Maltese, Mandarin, Marathi.

Language: polish

Unique tokens: 30

Tokens: ['a', 'b', 'c', 'd', 'e', 'f', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'ʃ', 'u', 'v', 'w', 'z', 'ɛ', 'g', 'i', 'j', 's', 'ʃ', 'z', 'ʒ']

Language: portuguese

Unique tokens: 29

Tokens: ['a', 'b', 'd', 'e', 'f', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 's', 't', 'u', 'v', 'w', 'z', 'ɛ', 'ɔ', 'ə', 'g', 'i', 'r', 'R', 'ʃ', 'u', 'ʒ']

Language: romanian

Unique tokens: 25

Tokens: ['a', 'b', 'd', 'd͡ʒ', 'e', 'f', 'h', 'i', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'ts', 'u', 'v', 'ɔ', 'g', 'i', 'j', 'ʒ']

Language: russian

Unique tokens: 45

Tokens: ['a', 'b', 'bʲ', 'd', 'e', 'i', 'j', 'k', 'kʲ', 'l', 'm', 'mʲ', 'n', 'nʲ', 'nʲ:', 'o', 'p', 'pʲ', 'r', 's', 'sʲ', 't', 'tʲ', 't͡ɕ', 'u', 'v', 'x', 'z', 'zʲ', 'ɛ', 'ɔ', 'ɛ:', 'ə', 'ɛ', 'g', 'gʲ', 'ɣ', 'i', 'ɪ', 'r', 's', 'ʉ', 'u', 'u', 'ʒ']

Language: spanish

Unique tokens: 25

Tokens: ['a', 'b', 'd', 'e', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'ʃ', 'u', 'w', 'x', 'ð', 'ŋ', 'g', 'j', 'β', 'θ']

Language: standard_arabic

Unique tokens: 31

Tokens: ['a', 'a:', 'b', 'd', 'd͡ʒ', 'e', 'f', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'q', 'r', 's', 't', 'u', 'w', 'x', 'z', 'ð', 'ħ', 'ɣ', 'ʃ', 'ʔ', 'ʕ', 'β', 'θ']

Language: standard_german

Unique tokens: 39

Tokens: ['a', 'a:', 'b', 'c', 'd', 'e:', 'f', 'h', 'i', 'i:', 'k', 'l', 'm', 'n', 'o', 'o:', 'p', 'r', 's', 't', 'ts', 'u', 'v', 'x', 'y:', 'z', 'ŋ', 'ɐ', 'ɑ', 'ɔ', 'ə', 'ɛ', 'g', 'ɪ', 'ɪ:', 'ʃ', 'ʊ', 'ʏ', 'ʔ']

Language: tamil

Unique tokens: 26

Tokens: ['a', 'a:', 'e', 'i', 'j', 'k', 'l', 'l:', 'm', 'm:', 'n', 'n:', 'o', 'p', 't', 't:', 'ʃ', 'u', 'ɪ', 'l:', 'ŋ', 'ŋ:', 'r', 'ɻ', 'u']

Language: turkish

Unique tokens: 28

Tokens: ['a', 'b', 'd', 'd͡ʒ', 'e', 'f', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 's', 't', 'ts', 'ʃ', 'u', 'v', 'y', 'z', 'z:', 'œ', 'g', 'w', 'r', 'ʃ']

Figure 10: Phoneme-token inventory of Polish, Portuguese, Romanian, Russian, Spanish, Standard Arabic, Standard German, Tamil, Turkish.

Language: twi_asante

Unique tokens: 24

Tokens: ['a', 'b', 'd', 'e', 'f', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'u', 'w', 'z', 'ɔ', 'ɛ', 'g', 'ʃ']

Language: ukrainian

Unique tokens: 31

Tokens: ['a', 'b', 'd', 'i', 'j', 'k', 'l', 'lʲ', 'm', 'n', 'nʲ', 'o', 'p', 'r', 'rʲ', 's', 't', 'tʲ', 'ts', 'ʃ', 'u', 'v', 'x', 'z', 'zʲ', 'ɔ', 'ɛ', 'h', 'ɪ', 'ʃ', 'ʒ']

Figure 11: Phoneme-token inventory of Asante Twi, Ukrainian.