Beyond *Bouba & Kiki*: Does Sound Symbolism scale across 27 Languages?

Anonymous ACL submission

Abstract

This paper investigates whether phonemes consistently convey size-related meaning across languages, a phenomenon known as sound symbolism. We compile a typologically diverse dataset of 810 adjectives (30 per language across 27 languages and 13 families), each phonemically transcribed and validated using native speaker recordings. Using bag-of-phoneme vectors and baseline classifiers, we show that size semantics can be predicted from phonological features with statistically significant accuracy, even across unrelated languages. Surprisingly, consonants such as /q/ and /ɦ/ emerge as highly predictive, challenging prior work that emphasizes vowel symbolism. To separate symbolic patterns from language-specific cues, we introduce an adversarial model that penalizes language prediction while preserving size-related information. Under the adversarial setup, however, classification accuracy drops to near chance—suggesting that much of the symbolic signal may be entangled with language-specific structure, or that larger datasets may be needed to detect more subtle cross-linguistic patterns.

1 Introduction

If you were watching a superhero movie called *Lamonians vs. Grataks: The Phoneme Accords*, chances are you'd already be rooting for the Lamonians because they *sound* like they would be nicer. In a 2009 *Guardian* article, linguist David Crystal posed a thought experiment: when asked to judge two fictional alien races, most people instinctively sided with the Lamonians, drawn to the soft consonants (/l/, /m/, /n/) and long vowels and diphthongs that give the name its gentle, likable tone (Crystal (2009)). This phenomenon in which specific sounds systematically convey particular meanings is known as *sound symbolism*, and it challenges the long-standing linguistic assumption that form and meaning are entirely arbitrary.

Sound symbolism is most familiar in onomatopoeia—words like *buzz* or *crash* that imitate real-world sounds. It also manifests systematically across languages: in Yucatec Maya, vowel length signals event duration (Guen, 2013); in Swedish, the prefix *pj-* marks pejoration (Åsa Abelin, 1999); and in Japanese, consonants in food mimetics reflect perceived crispness (Raevskiy et al., 2023). Beyond language, sound symbolism has commercial applications: high-frequency sounds like /i/, /e/, and /v/ are associated with luxury and increase consumer appeal in branding and hospitality (Motoki et al., 2023).

Despite evidence from individual languages, identifying cross-linguistic sound symbolism remains methodologically difficult. First, shared language ancestry makes it hard to disentangle universal patterns from inherited ones. Second, phonological inventories differ—some languages lack certain sounds—obscuring potential effects. Third, symbolic patterns must be separated from language-specific signals like dialectal variation, a task traditional methods often fail to resolve. Yet uncovering cross-linguistic sound symbolism has broader value. It may reveal cognitive universals in perception, improve cross-lingual transfer in low-resource NLP, and guide data-driven brand naming in commercial applications.

In this work, we investigate whether sound symbolism for size holds across typologically diverse languages by testing if adjectives meaning "small" and "large" consistently share phonological features, regardless of language family. Our approach includes a novel adversarial setup designed to isolate potentially universal sound-symbolic patterns while controlling for language-specific influences. Our contributions are:

- A cross-linguistic dataset of 800+ size adjectives (30 per language) from 27 languages across 13 language families, phonemically

transcribed in IPA and validated through native speaker recordings to capture contrastive phonological distinctions.

- Baseline classification experiments using logistic regression and decision trees on bag-of-phoneme vectors, with languages grouped into typological bins using Levenshtein distance (a measure of phonological similarity) to control for relatedness and assess cross-family generalization.

- An adversarial neural model with a phoneme encoder, a size classifier, and a language discriminator trained via minimax optimization to separate symbolic patterns from language-specific signals.

- Identification of predictive phonemes that correlate with size semantics across unrelated languages, revealing robust sound-symbolic patterns and the unexpected role of consonants.

2 Related Works

The idea that speech sounds might carry meaning isn't new, but Sapir (1929) was the first to test it experimentally. In his study, participants reliably judged nonsense words like *mal* to represent larger objects than *mil*, associating /a/ with largeness and /i/ with smallness. This challenged the assumption that word forms are entirely arbitrary. Since then, size symbolism has emerged as one of the most consistent patterns in sound symbolism. Yet most research remains focused on single languages, often Indo-European, and few computational approaches have explored whether such patterns generalize across diverse language families.

2.1 Linguistic Perspectives on Sound Symbolism

Although the idea that sounds can carry meaning is compelling, evidence from natural language remains uneven. Some studies suggest sound-meaning relationships are restricted to specific lexical categories or language families, while others demonstrate more universal cross-linguistic patterns that transcend typological boundaries.

Winter and Perlman (2021), for example, found that certain phonemes (such as /i/, /ɪ/, and /t/) were strongly associated with smallness in English-size adjectives, but this pattern did not generalize to nouns or other word types, suggesting that sound symbolism may be limited to specific lexical categories. Blasi et al. (2016) analyzed the core vocabulary of more than 6,000 languages and found phoneme–meaning associations (e.g., /i/ with "small", /l/ with "tongue") that emerged independently across language families.

Shape symbolism offers another compelling case: the maluma/takete effect, where soft-sounding pseudowords are matched to round shapes and sharp-sounding ones to spiky shapes, has been widely observed across cultures. Sidhu et al. (2021) extended this to English nouns, showing that round-object words tend to include /m/ and /u/, while spiky-object words more often include /k/ and /t/ suggesting that the sound of a word may reflect aspects of its meaning.

2.2 Computational Approaches to Sound Symbolism

Unlike linguistic studies that rely on curated word lists and controlled tasks, recent computational work explores whether models trained on large-scale data internalize symbolic patterns like the Bouba–Kiki effect, where people intuitively match soft sounds to round shapes and sharp sounds to spiky ones. Passi and Arun (2024) showed that this effect holds even for reversed speech, pure tones, and environmental sounds, and is best predicted by sound frequency—challenging the idea that it depends on mouth shape.

Alper and Averbuch-Elor (2023) found that CLIP and Stable Diffusion reliably associate pseudowords with matching visual features, suggesting that vision–language models can learn sound–meaning mappings from text–image co-occurrence alone. Similarly, Loakman et al. (2024) found that larger LLMs and VLMs, such as GPT-4, moderately align with human judgments on size symbolism—despite having no explicit access to phonetic information. Interestingly, even fictional names reveal learnable sound–meaning patterns across languages. Kilpatrick et al. (2023) trained random forests on Pokémon names in Japanese, Chinese, and Korean to classify them as pre- or post-evolution, finding that models outperformed human participants and relied on phonemes like /d/, /g/, and long vowels—demonstrating that symbolic cues like size and strength generalize across both natural and invented lexicons.

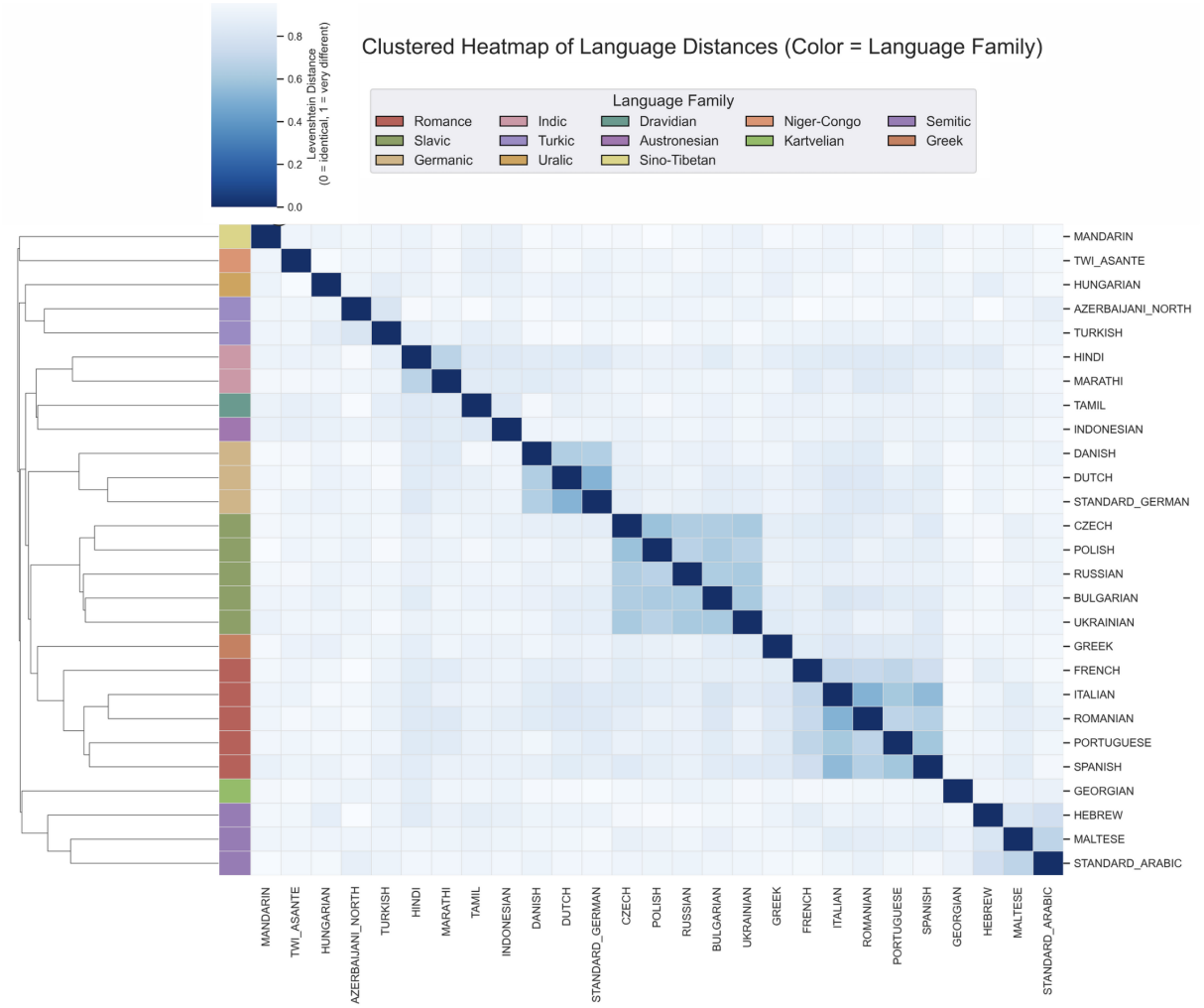

Figure 1: Levenshtein distance heatmap showing pairwise phonological similarity between the 27 languages in our dataset. Languages are grouped hierarchically and colored bars indicate language families.

3 Methodology

3.1 Data Collection

To investigate sound symbolism across typologically diverse languages, we compiled a dataset of 810 adjectives—30 per language (15 denoting smallness, e.g., tiny, minuscule; 15 denoting largeness, e.g., huge, enormous)—spanning 27 languages from 13 language families. Languages were selected using the World Atlas of Language Structures (WALS) to maximize both genetic and areal diversity, while ensuring feasibility in terms of speaker access and resource availability.

Each adjective list was constructed by translating English seed words using DeepL, Google Translate, and bilingual dictionaries. We manually filtered out borrowings and transliterations to prioritize native lexical items. Words were then transcribed into the International Phonetic Alphabet (IPA) using phonemic transcription, which captures contrastive sound units while abstracting away from fine-grained phonetic variation. This choice was motivated by both practical and linguistic considerations. Phonemic representations provide greater consistency across typologically diverse languages and reduce confounds from dialectal variation, particularly in languages such as Arabic and Spanish. Focusing on phoneme-level distinctions allowed us to isolate symbolic features most relevant to meaning while preserving cross-linguistic comparability.

We initially experimented with large language models such as GPT and Claude for automatic transcription, but found that their outputs often deviated from canonical phonology, even in well-resourced languages. To ensure accuracy, we adopted a hybrid transcription pipeline. Where available, we used the XPF corpus—a rule-based, linguistically grounded phoneme-mapping tool—to generate initial IPA forms (Priva et al., 2021). We then col-

lected audio recordings of the adjective lists from native speakers via language exchange platforms such as Tandem. These recordings were not used as model input, but served as references for transcription correction.

A trained linguist and phonetician reviewed the recordings using acoustic analysis tools and manually revised or reconstructed each transcription. In cases where no initial transcription was available, phonological forms were created from scratch based on the recordings. Domain experts, including language instructors and researchers, also reviewed or annotated the data for several languages to ensure quality and accuracy.

To control for shared linguistic ancestry, we used a precomputed dataset of normalized Levenshtein distances between languages, compiled by Svend V. Nielsen (Nielsen, 2017) following the method introduced by Bakker et al. (2009). These distances (Figure 1) were derived from phonologically transcribed Swadesh lists—a standardized set of core vocabulary items (eg. "I," "water" etc). Nielsen's implementation used a reduced 40-word version of the list, chosen for high cross-linguistic comparability. The resulting values reflect the average number of phonological edits needed to align word pairs across languages. We used these distances to group languages into similarity bins for our classification experiments.

While the per-language dataset is relatively small, collecting speaker-validated, phonemically transcribed data across 27 languages poses substantial logistical challenges. Our primary goal was to ensure lexical balance and phonological comparability, even at a modest scale. To compensate for this, particularly in our adversarial setup, we pretrained the phonological encoder on over three million word–pronunciation pairs extracted using WikiPron[1] (Lee et al., 2020), an open-source tool for systematically retrieving IPA transcriptions from Wiktionary. This large-scale pretraining enables the encoder to learn general phonotactic and articulatory patterns across a broad range of languages, enhancing its ability to detect symbolic structure in our more controlled downstream dataset.

3.2 Baseline Classifiers

We first evaluated whether size-related sound-symbolic patterns could be detected using simple, interpretable models. To this end, we trained logistic regression and decision tree classifiers to predict size semantics from phonemic input alone.

Each word was represented as a bag-of-phoneme vector based on IPA symbol counts. To ensure consistency across transcription styles, we limited features to core IPA characters, excluding diacritics and suprasegmentals.

To test cross-linguistic generalization, we grouped training languages into three bins of typological similarity relative to each target language—Most Similar, Somewhat Similar, and Least Similar—based on the Levenshtein distance matrix described above. This allowed us to examine whether classifiers trained on typologically related languages performed better than those trained on distant ones.

Logistic regression used default regularization. Decision trees were tuned for maximum depth (3–15) and minimum samples per split (2–10) to control overfitting. Accuracy was used as the primary evaluation metric.

To analyze feature importance, we extracted and aggregated logistic regression coefficients across models and bins to identify phonemes most strongly associated with smallness or largeness. We did not interpret decision tree feature importances due to structural variability across runs.

3.3 Adversarial Scrubber

Our goal is to learn phonological representations that retain sound-symbolic information (e.g., phonetic cues relevant to size) while suppressing language-specific patterns that could confound cross-linguistic generalization. We adopt a two-stage approach that combines phonetic pretraining with adversarial representation learning. The model is optimized to predict semantic size (small vs. large) while minimizing its ability to recover the language family from the same embedding.

In the first stage, we pretrain a compact BERT encoder on 1.6 million IPA-transcribed words from WikiPron using a masked language modeling objective. This step encourages the model to capture general phonological structure across typologically diverse languages. The encoder consists of 2 transformer layers with 128 hidden units and a vocabulary of 80 IPA characters plus standard special tokens. After 2 epochs of training, the encoder is frozen and used as a fixed feature extractor.

We freeze our BERT model and then fine-tune a linear embedding layer, a sound symbolism clas-

[1]Available under the Apache 2.0 license.

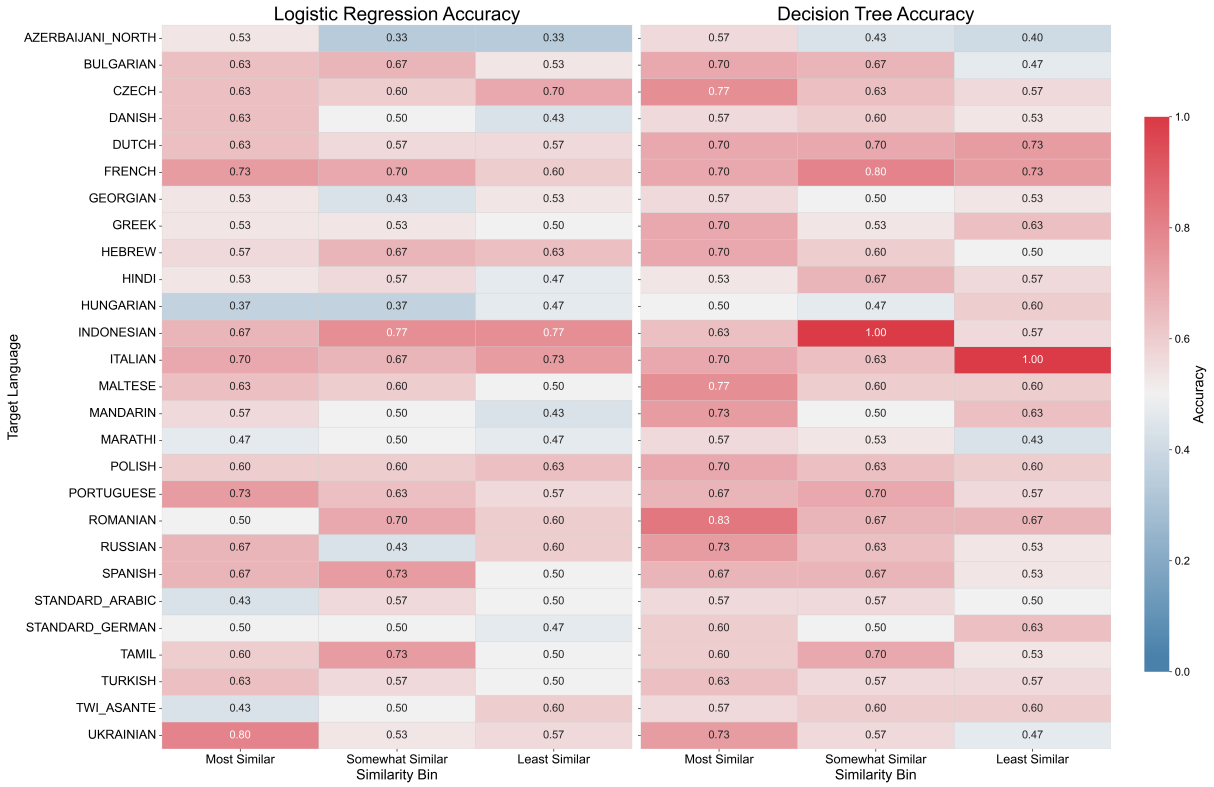

Figure 2: Detailed accuracy heatmap by target language and similarity bin. Each cell shows classification accuracy for logistic regression (left) and decision tree (right) models.

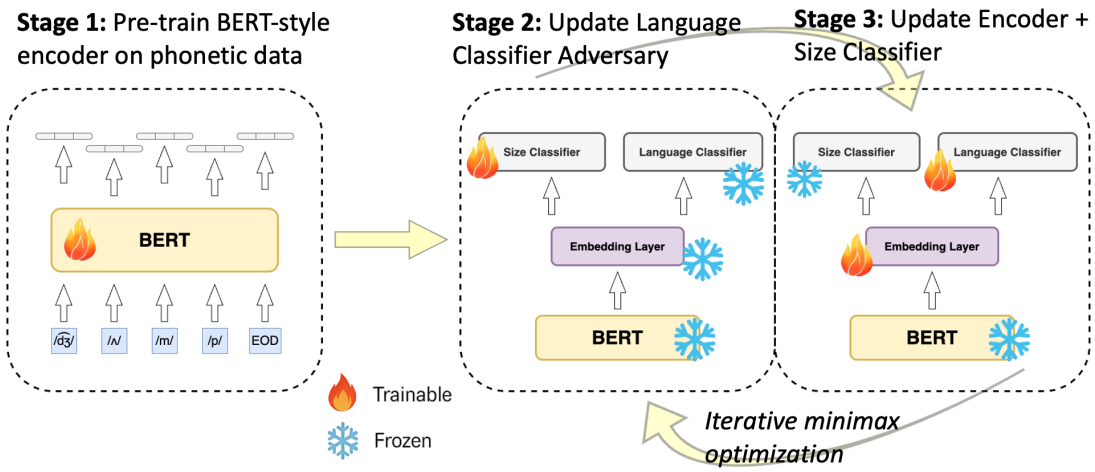

Figure 3: Overview of our adversarial training pipeline. Stage 1: Pretrain BERT on IPA-transcribed words using masked language modeling. Stage 2: Train a language classifier on frozen embeddings. Stage 3: Fine-tune the embedding layer and symbolism classifier while suppressing language-specific cues via adversarial loss.

sifier, and a language adversary using our curated dataset, in which each word is annotated with both a binary size label and a language family label. The model architecture includes the following components:

- **Encoder (E):** A linear embedding layer that takes the BERT embeddings as input.
- **Sound Symbolism Classifier (C):** A single linear layer followed by a softmax over two classes (small vs. large).
- **Language Adversary (A):** A multi-class classifier trained to predict the word's language family from the same embedding.

To encourage the encoder to discard language-specific information, we optimize a minimax objective:

$$\min_{E,C} \max_{A} \quad \mathcal{L}_{\text{symbolism}}(C(E(x)),\ y_{\text{size}}) - \lambda \cdot \mathcal{L}_{\text{language}}(A(E(x)),\ y_{\text{lang}}) \tag{1}$$

Here, $\mathcal{L}_{\text{symbolism}}$ and $\mathcal{L}_{\text{language}}$ are cross-entropy losses for size and language classification, respectively. We used a conservative base value of $\lambda = 0.01$, with annealing beginning after epoch 10 to prevent premature over-scrubbing. Specifically, λ increased over training using the schedule $\lambda_t = 0.01 \cdot (1 - 0.95^{t-9})$, allowing the encoder to first learn phonological structure before gradually suppressing language-specific patterns.

Training alternates between two steps within each batch: (1) updating the adversary to improve language prediction, and (2) updating the encoder and symbolism classifier to improve size prediction while reducing the adversary's ability to recover language identity. During the second step, the adversary is frozen, and the encoder is encouraged to produce language-invariant representations.

We project BERT embeddings into a 32-dimensional space and apply dropout ($p = 0.4$) before classification. Language identity was represented using a coarse categorical label (0, 1, 2), based on a simplified grouping created during preprocessing. These categories roughly reflected genealogical relationships—for example, grouping Romance languages together—but also included some typologically diverse combinations, such as Hindi, Russian, and German. This decision was motivated by our limited dataset size, which made fine-grained modeling of language structure impractical. While these groupings do not capture the full complexity of linguistic relationships, they offered a manageable way to test whether the model could suppress broad language-level patterns while still retaining symbolic cues relevant to size.

4 Results

4.1 Baseline Classification Results

Our classification experiments show that phonological features carry predictive information about size semantics across languages. Both logistic regression and decision tree classifiers performed consistently above the 50% random baseline across all three similarity bins.

Model	Most Similar	Somewhat Similar	Least Similar
Logistic Reg.	59.0%	57.3%	54.4%
Decision Tree	65.6%	61.7%	58.1%

Table 1: Mean accuracy (%) by model and similarity bin.

All reported results were statistically significant compared to chance ($p < 0.001$ for all bins with decision trees; $p < 0.001$, 0.01, and 0.05 for logistic regression depending on the bin). A one-way ANOVA revealed that differences in performance across bins were not statistically significant for either model ($p = 0.27$ for logistic regression; $p = 0.21$ for decision trees).

Figure **??** shows the distribution of model performance across similarity bins, and Figure 2 provides a breakdown by target language.

4.2 Phonological Features & Size Symbolism

Analysis of feature importance in our models revealed several phonemes that consistently predict size semantics across languages. Contrary to expectations from prior literature focusing primarily on vowel symbolism, our baseline results suggest that consonants may also play a significant role in sound-symbolic associations.

Based on averaged logistic regression coefficients across all languages, the phonemes most strongly associated with largeness include /ɦ/, /ɜ/ and /ŋ/. For smallness, the most predictive phonemes include /q/, /ʔ/ and /ð/ (figure 4).

These results complicate the predictions of the Frequency Code hypothesis (CITE), which suggests that lower-frequency sounds should iconically convey largeness. While some findings align with this view—for instance, the association of the back vowel /o/ with largeness—others are less expected. A number of the top predictors are consonants, including some that are relatively rare across languages, suggesting that sound-symbolic patterns may not be limited to the well-attested vowel contrasts. Instead, they may also reflect more language-specific or articulatory factors.

4.3 Adversarial Training Results

To evaluate whether language-specific signals could be suppressed without eliminating sound-symbolic structure, we trained an adversarial model using the best-performing hyperparameters identified through pilot experiments that varied embed-

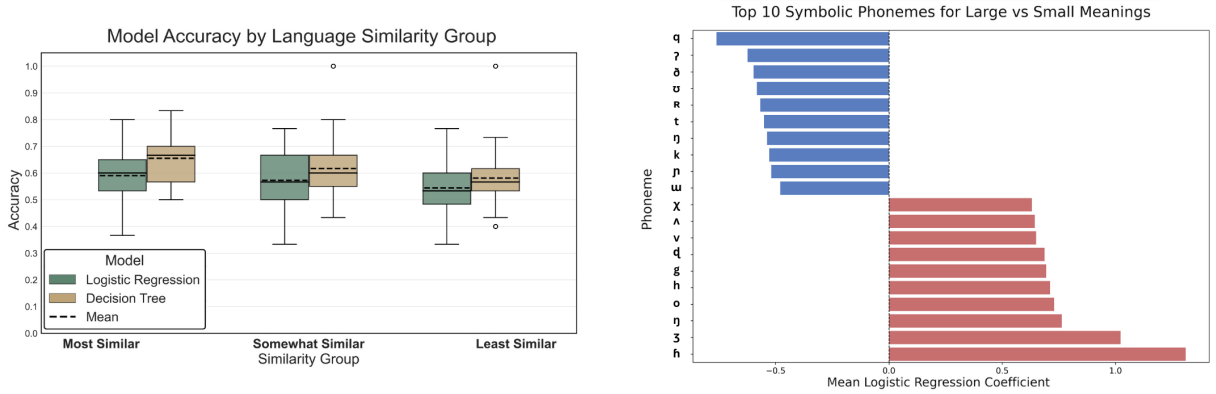

Figure 4: **Left:** Model accuracy across language similarity groups (logistic regression and decision tree classifiers). **Right:** Most predictive phonemes for size classification using logistic regression.

ding dimensionality, dropout rate, and adversarial loss weight. These experiments aimed to balance effective language suppression with preservation of symbolic information. The final model achieved 52.5% symbolic classification accuracy and 22.2% language classification accuracy on the test set. While this indicates partial success in reducing language cues, symbolic performance remained below that of baseline classifiers.

The model showed a clear bias toward predicting largeness, with a recall of 0.78 for "large" words compared to just 0.27 for "small." In other words, the model frequently mislabeled small words as large. This asymmetry may reflect that largeness cues were more robustly encoded across languages and better preserved under adversarial pressure. In contrast, phonological patterns linked to smallness may have been subtler or more language-specific, making them more susceptible to being scrubbed during training.

4.4 Ablation Experiments

Table 2 summarizes the results of our ablation experiments. The experiment ID (leftmost column) begins at 0 for the baseline model.

1. **Scrambled Size Labels:** To verify if classifiers learned symbolic structures instead of memorizing artifacts, we did a control experiment with scrambled size labels. For each language, we randomly shuffled test set size labels while keeping class balance. This checked if models truly captured phoneme–meaning associations, expecting chance performance if semantics were random. Logistic regression didn't exceed chance in any similarity bin ($p = 0.82$, 0.29, and 0.20 for most, somewhat, and least similar bins). Decision trees, however, worked above chance in all bins ($p < 0.001$, 0.0064, 0.0048).

2. **Vowels only:** Our ablation was driven by an unexpected finding: consonants, not vowels, were most predictive of size in our baseline logistic regression. This is contrary to previous sound symbolism research which emphasizes vowels' roles, particularly high front vowels for smallness and back or low vowels for largeness. Despite using only vowels, both logistic regression and decision tree classifiers showed strong performance with minimal decline. These findings imply vowels hold a symbolic significance but also suggest consonants play an equally important role in size semantics across languages. Further research with larger datasets could clarify the roles of different phoneme classes.

3. **High Frequency Phones only:** This ablation tested whether performance would change when models are restricted to only the most common phonemes cross-linguistically. The list of 23 segments was provided by our collaborating phonetician and includes phones that are widely attested across the world's languages. The goal was to test if symbolic patterns persist using only globally common sounds. Both classifiers performed nearly as well as the full model. This suggests that much of the sound-symbolic signal can be recovered using only the most typologically frequent phones, and that rarer or language-

ID	Ablation Condition	LR Mean	DT Mean	LR / DT by Bin		
0	Baseline (all phonemes)	56.9	61.8	59.0 / 57.3 / 54.4	|	65.6 / 61.7 / 58.1
1	Scrambled Labels	51.7	56.5	52.5 / 51.1 / 51.4	|	56.4 / 57.0 / 57.0
2	Only Vowels	57.0	61.0	55.9 / 56.4 / 58.5	|	60.6 / 60.6 / 61.7
3	High-Frequency Phones (see Appendix A)	57.9	62.1	60.4 / 54.7 / 57.9	|	67.3 / 60.6 / 62.1
4	Only Plosives (see Appendix A)	54.5	58.4	56.2 / 52.7 / 54.6	|	59.5 / 60.5 / 55.2
5	Only Nasals (see Appendix A)	51.2	54.2	51.5 / 51.5 / 50.6	|	53.6 / 54.7 / 54.3

Table 2: Mean and 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 are listed in Appendix A.

specific phonemes may not be necessary to detect size-related patterns.

4. **Consonants only (Stops vs. Nasals):** To identify consonant types that contribute to sound-symbolic patterns, we analyzed stops and nasals. These were chosen for their high frequency across languages and natural contrast in articulation: stops involve closure and bursts, while nasals are sonorants produced using continuous nasal airflow. Stops (e.g., /p/, /t/, /k/) are often associated with impact, abruptness, or emphasis in sound-symbolic research, whereas nasals (e.g., /m/, /n/, /ŋ/) tend to have a more resonant and vowel-like acoustic profile. Stops outperformed nasals in both models: 54.5% vs. 51.2% in logistic regression and 58.4% vs. 54.2% in decision trees. Both exceeded the scrambled-label baseline but were less accurate than the vowel-only condition, despite consonants being top-weighted features in the baseline.

5 Discussion

The success of logistic regression and decision tree classifiers in predicting size semantics indicates that phonological forms encode systematic, cross-linguistic sound-symbolic patterns. Statistical tests showed classification accuracy was significantly above chance, especially for decision trees ($p < 0.001$). Notably, performance stayed above chance even in the least similar bin where training and test languages were typologically distant. The consistent accuracy across bins suggests phonological features related to size symbolism aren't restricted by linguistic ancestry. This challenges the idea that sound symbolism is only from shared linguistic ancestry, suggesting a broader cognitive basis for phoneme–meaning associations across diverse languages. Decision trees generally outperformed logistic regression, especially in similar bins, but with possible overfitting. Logistic regression offers a more conservative and stable estimate, consistently capturing robust cross-linguistic symbolic patterns. Logistic regression and decision tree models unexpectedly outperformed the adversarial approach, highlighting the challenge of removing language-specific data while preserving necessary signals for classification. Phonological cues for size might overlap with language patterns, meaning suppressing one can also suppress the other. Baseline models did not have this restriction and used all available variations, including language-specific ones. That said, the ablations suggest no single phoneme group—like consonants—consistently drives performance. And while the baseline results seem promising, the adversarial setup shows that much of the signal may not reflect true cross-linguistic symbolism.

6 Conclusion

In this work, we showed that simple, interpretable models trained on bag-of-phoneme features can predict "small" vs. "large" adjectives across 27 diverse languages at rates above chance (e.g., logistic regression at 59–54% accuracy and decision trees up to 66%). An adversarial framework reduced language-identification to 22% while retaining 52.5% size-prediction accuracy, highlighting promising future potential for a more detailed adversarial approach to modeling sound symbolism. Our feature-weight analysis confirmed the role of vowels and revealed consonantal predictors (e.g., /ɦ/, /q/), challenging vowel-centric accounts. Moving forward, expanding the dataset's scale and phonetic granularity—and refining adversarial objectives—will be key to isolating truly universal sound-symbolic patterns.

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 with limited academic or online presence—posed a significant logistical challenge. 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. While we aimed for genealogical diversity using WALS and intentionally 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.

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)*.

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 lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13(1):169–181.

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.

David Crystal. 2009. The ugliest words; marilyn monroe's dog; and wacko races. *The Guardian*, July 17, 2009. Accessed May 18, 2025.

Olivier Le Guen. 2013. Ideophones in yucatec maya. Manuscript.

Alexander James Kilpatrick, Aleksandra Ćwiek, and Shigeto Kawahara. 2023. Random forests, sound symbolism and pokémon evolution. *PLOS ONE*, 18(1):e0279350.

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.

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.

Kosuke Motoki, Jaewoo Park, Abhishek Pathak, and Charles Spence. 2023. Creating luxury brand names in the hospitality and tourism sector: The role of sound symbolism in destination branding. *Journal of Destination Marketing & Management*, 30:100815.

Svend V. Nielsen. 2017. Mds plot of the world's languages. Accessed: 2025-07-25.

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.

Uriel Cohen Priva, Emily Strand, Shiying 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.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12:225–239.

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.

Bodo Winter and Marcus Perlman. 2021. Size sound symbolism in the english lexicon. *Glossa: a journal of general linguistics*, 6(1):79.

Åsa Abelin. 1999. *Studies in Sound Symbolism*. Ph.D. thesis, Göteborg University. Doctoral Dissertation.

A Appendix A: Ablation Phone Sets

Ablation experiment index	Experiment Name	Phonemes considered
0 and 1	Baseline and Scrambled labels	/p/, /b/, /t/, /d/, /ʈ/, /ɖ/, /c/, /ɟ/, /k/, /g/, /q/, /ɢ/, /ʔ/, /m/, /ɱ/, /n/, /ɳ/, /ɲ/, /ŋ/, /ɴ/, /ʙ/, /r/, /ʀ/,/ⱱ/, /ɾ/, /ɽ/, /ɸ/, /β/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /ʂ/, /ʐ/, /ç/, /ʝ/, /x/, /ɣ/, /χ/, /ʁ/, /ħ/, /ʕ/, /ʜ/, /ʢ/, /ǀ/, /ǃ/, /ǂ/, /ǁ/, /ɓ/, /ɗ/, /ʄ/, /ɠ/, /ʛ/, /pʼ/, /tʼ/, /kʼ/, /sʼ/, /ʍ/, /w/, /ɥ/, /ʜ/, /ʢ/, /ʡ/, /ʧ/, /ʤ/, /i/, /y/, /ɨ/, /ʉ/, /ɯ/, /u/, /ɪ/, /ʏ/, /ʊ/, /e/, /ø/, /ɘ/, /ɵ/, /ɤ/, /o/, /ə/, /ɛ/, /œ/, /ɜ/, /ɞ/, /ʌ/, /ɔ/, /æ/, /a/, /ɶ/, /ɑ/, /ɒ/
2	Only vowels	/i/, /y/, /ɨ/, /ʉ/, /ɯ/, /u/, /ɪ/, /ʏ/, /ʊ/, /e/, /ø/, /ɘ/, /ɵ/, /ɤ/, /o/, /ə/, /ɛ/, /œ/, /ɜ/, /ɞ/, /ʌ/, /ɔ/, /æ/, /a/, /ɶ/, /ɑ/, /ɒ/
3	High Frequency phones	/p/, /t/, /k/, /g/, /ʔ/, /b/, /d/, /m/, /n/, /ɲ/, /f/, /v/, /s/, /z/, /h/, /j/, /l/, /w/, /i/, /e/, /o/, /a/, /u/
4	Plosives	/p/, /b/, /t/, /d/, /ʈ/, /ɖ/, /c/, /ɟ/, /k/, /g/, /q/, /ɢ/, /ʔ/
5	Nasals	/m/, /ɱ/, /n/, /ɳ/, /ɲ/, /ŋ/, /ɴ/