Which phonological features express size sound symbolism? – A cross-linguistic analysis with extreme gradient machine learning

Melissa Ebert, Alexander Kilpatrick & Aleksandra Ćwiek

(Humboldt University of Berlin, Nagoya University of Business and Commerce & Leibniz-Centre General Linguistics, Berlin)

Keywords: Size sound symbolism, phonological features, dimensional antonym adjective pairs, XGBoost, machine learning

Research on size sound symbolism has revealed patterns indicating that certain phonemes appear to express certain size attributes. For example, cross-linguistically smallness is typically associated with high front vowels and largeness with low back vowels (e.g., Sapir 1929). Studies have shown that machine learning algorithms can understand size sound symbolism if they are trained on phonemes of words (e.g., Winter & Perlman 2021). This study uses a machine learning approach to investigate if phonological features, rather than phonemes, that indicate largeness or smallness across languages can be identified. As phonemic inventories differ across languages, analysing phonological features might uncover further underlying patterns.

The data and analysis scripts are available in the OSF repository: https://osf.io/smbt8/?view_only=d7c96d8b966f4577afa5403a110a22d1. The dataset consists of 11 antonym adjective pairs in 22 languages (7 language families) in which each pair entails a dimensionally large and small adjective, e.g., *big-small*, *thick-thin*, *long-short*, *far-near* (Haynie et al. 2014, Fuchs et al. 2019). We applied a modified set of 26 phonological features from the PanPhon database (Mortensen et al. 2016) to the adjectives, utilizing the same features for both vowels and consonants (Odden 2005). We then trained extreme gradient machine learning algorithms to classify the adjectives into binary (large/small dimension) categories. The data was processed in R (R Core Team 2022) and the algorithms were constructed using the XGBoost (Chen & Guestrin 2016) and caret (Kuhn 2008) packages. The dependent variable for the algorithms was the binary dimension classification (large/small) and the independent variables were counts of phonological features divided by total number of phonemes in each sample to mitigate the effect of word length. Given the sample size (N = 502), we used K-fold cross-validation (K = 5) and examined algorithm behavior using feature importance and distribution of features to large/small samples.

The algorithms accurately classified 82.4% of the samples (*SD* = 4.6%). In all cases, classification accuracy was significant ($p < 0.001$). Table 1 shows the feature importance scores and distribution of the ten most important phonological features for the large/small dimension classification.

The high classification accuracy indicates that dimensional size opposition is robustly expressed sound-symbolically and that different languages express it through similar phonological features. The feature importance scores partly resonate with established research (e.g., coronal, anterior, and front pointing towards smallness) but also exhibit new associations that need to be investigated further (e.g., continuant being the most important feature and indicating smallness, or back pointing towards smallness).

Feature	Importance	Distribution
Continuant	95.88	Small
Voiced	87.95	Large
Minus low	86.47	Small
Coronal	83.02	Small
Back	82.03	Small
Anterior	81.72	Small
Front	68.42	Small
Minus high	65.71	Large
Obstruent	61.99	Large
Sonorant	61.13	Small

Table 1: The ten most important features in the algorithm, their feature importance scores, and distribution to the large/small dimension classification.

References

Chen, Tianqi & Carlos Guestrin (2016), XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM, 785–794.

Fuchs, Susanne, Egor Savin, Stephanie Solt, Cornelia Ebert & Manfred Krifka (2019), Antonym adjective pairs and prosodic iconicity: evidence from letter replications in an English blogger corpus, *Linguistics Vanguard* 5(1), 20180017.

Haynie, Hannah, Claire Bowern & Hannah LaPalombara (2014), Sound symbolism in the languages of Australia, *PLoS ONE* 9(4), e92852.

Kuhn, Max (2008), Building Predictive Models in R Using the caret Package, *Journal of Statistical Software* 28(5), 1–26.

Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer & Lori Levin (2016), PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 3475–3484.

Odden, David (2005), *Introducing Phonology*, Cambridge: Cambridge University Press.

R Core Team (2022), R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria.

Sapir, Edward (1929), A study in phonetic symbolism, *Journal of Experimental Psychology* 12(3), 225–239.

Winter, Bodo & Marcus Perlman (2021), Size sound symbolism in the English lexicon, *Glossa: a journal of general linguistics* 6(1), 1–13.