# A Layered Bridge from Sound to Meaning: Investigating Cross-linguistic Phonosemantic Correspondences

**Andrea Gregor de Varda (andreagregor.devarda@studenti.unitn.it)**
CIMeC – Center for Mind/Brain Sciences, 31 Corso Bettini
Rovereto, TN 38068 IT

**Carlo Strapparava (strappa@fbk.eu)**
FBK – Fondazione Bruno Kessler, 18 Via Sommarive
Povo, TN 38123 IT

**Abstract**

The present paper addresses the study of cross-linguistic phonosemantic correspondences within a deep learning framework. An LSTM-based Recurrent Neural Network is trained to associate the phonetic representation of a word, encoded as a sequence of feature vectors, to its corresponding semantic representation in a multilingual and cross-family vector space. The processing network is then tested, without further training, in a language that does not appear in the training set and belongs to a different language family. The performance of the model is evaluated through a comparison with a monolingual and mono-family upper bound and a randomized baseline. After the assessment of the network's performance, the distribution of phonosemantic properties in the lexicon is inspected in relation to different (psycho)linguistic variables, showing a link between lexical non-arbitrariness and semantic, syntactic, pragmatic, and developmental factors.

**Keywords:** Phonosymbolism; cross-lingualism; deep learning

## Introduction

The relation between sound and meaning has held a particular fascination over philosophers and linguists since time immemorial. In the history of Western thought, the first documented inquiry dates back to the platonic dialogue *Cratylus*, where Socrates initially suggests that words fit their referents in virtue of the sounds they are made of. Other endeavors from ancient times that attempted to lead the semantic properties of referents back to their phonetic realization are found in the Hinduist (*Aitareya Aranyaka* III.2.6.2) and Buddhist (Kūkai's *Shôjijissôgi*) heritage. In recent times, this fascinating hypothesis has progressively lost the interest of scholars, especially in the structuralist linguistic tradition, which emphasized the arbitrariness in such relation (Saussure, 1964). The topic has recaptured its original attractiveness in the field of cognitive science, which has included in its domain of study topics whose inquiry traditionally fell under the wing of philosophy. Within this framework, the attention has particularly focused on the link between sound and shape. A prominent example of the naturally biased mappings came from Köhler's finding that, when asked to match two novel shapes with the non-words 'maluma' and 'takete', English-speaking adults tended to label as 'maluma' the curled shape, and as 'takete' the sharp one (Köhler, 1929). This germinal study paved the way to a number of replications and expansions of its findings, that reproduced Köhler's results in different geocultural contexts (Bremner et al., 2013) and at different developmental stages (Maurer et al., 2006). Since then, different studies have tackled the topic of non-arbitrariness in language from a broader perspective, showing that adults can associate visually presented characters (Koriat & Levy, 1977) and auditorily presented words (Berlin, 1995) of a foreign language to their meaning, with an accuracy above chance. Furthermore, it has been shown that participants perform above chance when pairing up words with opposite meanings in languages to which they have not been exposed (Nuckolls, 1999), and when estimating the concreteness of words from languages unknown to them (Reilly et al., 2017).

Recently, the notion of a sound-symbolic mapping between phonetic and semantic representations has gone from being a marginal – although appealing – matter to being integrated into broader theories of language evolution (Ramachandran & Hubbard, 2001), processing (Lockwood & Tuomainen, 2015) and acquisition (Asano et al., 2015; Imai et al., 2008). Indeed, rejecting the assumption of a totally arbitrary mapping between sound and meaning sensibly reduces the problem space of language emergence, establishing constraints on the consensus on word choice. Furthermore, a systematic relation between a sound and its referent might help with memory consolidation in the process of language acquisition (Sathian & Ramachandran, 2019). Phonosemantic correspondences have been shown to affect different cognitive faculties other than language, such as memory (Ramachandran & Hubbard, 2001), categorization (Lupyan & Casasanto, 2015), and emotion recognition (Slavova et al., 2019); moreover, they exert an influence on actional processes such as phonatory behavior (Parise & Pavani, 2011), spatial navigation (Rabaglia et al., 2016), and hand grip (Vainio et al., 2013). Given their widespread effects, it is reasonable to suspect that phonosemantic biases might not be limited to few circumscribed phonetic or semantic clusters, but may instead pervade the lexicon beyond the aforementioned anecdotal cases.

Within the computational framework, the analysis of phonosemantic biases has mainly followed two general trends (Gutiérrez et al., 2016): a localist approach, aimed at identifying some islands of non-arbitrariness in language (Sagi & Otis, 2008; Abramova et al., 2013; Abramova & Fernández, 2016), and a global program, directed toward an assessment of its pervasiveness and sistematicity (Shillcock et al., 2001; Monaghan et al., 2014; Tamariz, 2008; Dautriche et al., 2017). The first part of our work fits into the second trend, and aims to extend

the previous findings through an exploration of phonosemantic regularities beyond the limits of a single language. To our knowledge, few studies have tackled the topic of sound symbolism from a cross-linguistic perspective, generally focusing on a small set of concepts or words on a massively multilingual scale (Blasi et al., 2016; Wichmann et al., 2010; Johanssohn et al., 2020). Our study, in contrast, aims to perform a lexicon-wide analysis on a selected set of languages.

In the present work, we evaluated the performance of a Long Short-Term Memory network (LSTM) in associating phonetic vector sequences with semantic vectors in a multilingual space, reporting an above-chance score of the model in an unseen language. We constructed and compared three different models, based on the same neural architecture but characterized by different linguistic distances between the items in the training and in the test set. A cross-family model was trained in seven languages belonging to seven language families, and tested in a language that did not appear in the training set and corresponded to a different family. The performance of the cross-family model was compared with the results of (a) a mono-family model, trained and tested on eight Indoeuropean languages, and (b) a monolingual model, trained and tested on different subsets of the Italian lexicon. Our multilingual experiments were configured as zero-shot transfer tests, where the internal representations learned by the models were applied without further training to unseen vocabularies. The performances of the three networks were contrasted with their randomized counterparts, showing that the LSTMs learned a generative process where the semantic representation produced in response to a word's sound resembled the word's actual meaning more than it would be expected by chance. Then, we proposed an attempt to bridge the gap between localist and globalist approaches to phonosymbolism exploiting the LSTM predictions to derive a metric of a word's phonosemantic transparency. Finally, we adopted that metric to inspect the relation between the induced degree of non-arbitrariness and different (psycho)linguistic factors.

## Methods

In the present study, an LSTM-based Recurrent Neural Network was trained to associate the phonetic to the corresponding semantic representation of a word. Semantic representations consisted in 300-dimensional word embeddings in a multilingual vector space, whereas their corresponding phonetic features were expressed as sequences of phonetic vectors in 22 dimensions. The experimental pipeline is summarized in the flowchart in Figure 1.

### Semantic vectors

The semantic representations included in the model, provided by Facebook Research, consisted in multilingual word embeddings, generated with `fastText` from Wikipedia data (Bojanowski et al., 2017) and aligned in a common vector space with the RCSLS method (Joulin et al., 2018)[1]. The present study was conducted on word embeddings from fifteen languages, belonging to different language families and combined according to the experimental condition.

| Set | Language | Family |
|---|---|---|
| Training (cross-family condition) | Arabic | Afroasiatic |
| | Hungarian | Uralic |
| | Indonesian | Austronesian |
| | Thai | Tai-Kadai |
| | Vietnamese | Austroasiatic |
| | Turkish | Turkic |
| | Tamil | Dravidian |
| Training (mono-family condition) | Bengali | Indoeuropean |
| | Hindi | Indoeuropean |
| | Polish | Indoeuropean |
| | Ukrainian | Indoeuropean |
| | Dutch | Indoeuropean |
| | French | Indoeuropean |
| | Spanish | Indoeuropean |
| Training (monolingual condition) | Italian | Indoeuropean |
| Test | Italian | Indoeuropean |

Table 1: Languages and relative language families by experimental condition

### Phonetic vectors

For each word in the embedding dataset, we obtained its phonemic transcription with `Epitran`, a Python library for transliterating orthographic text in the International Phonetic Alphabet (IPA) format. Then, we converted the IPA string into a sequence of feature vectors with `PanPhon`, a package that traduces IPA segments into subsegmental articulatory features (Mortensen et al., 2016). We agree with Jakobson & Waugh (2011) when they assert that "most objections to the search for the inner significance of speech sounds arose because the latter were not dissected into their ultimate constituents" (p. 182). Hence, we chose not to directly hot-encode the IPA strings in order to allow the network to exploit the underlying similarities that make different phones more or less related to each other. For instance, [p] and [b] are similar in that they only differ in the feature [+/– voiced], whereas [t] and [u] differ by 13 subsegmental features. These internal asymmetries would have been lost with a raw hot-encoding over the IPA vocabulary. Before being loaded into the LSTM model, all the sequences were padded, with a maximum length of 29.

### Neural architecture

An LSTM-based Recurrent Neural Network was trained to map the sequences of phonetic feature vectors in input into the semantic vectors in output, with a many-to-one topological structure. The model was built with `Keras`, a deep learning framework for Python (Chollet et al., 2015); it included a masking layer, followed by a single LSTM layer and a dense layer. The number of hidden units of the masking and the

[1]Publicly available at `https://fasttext.cc/docs/en/aligned-vectors.html`

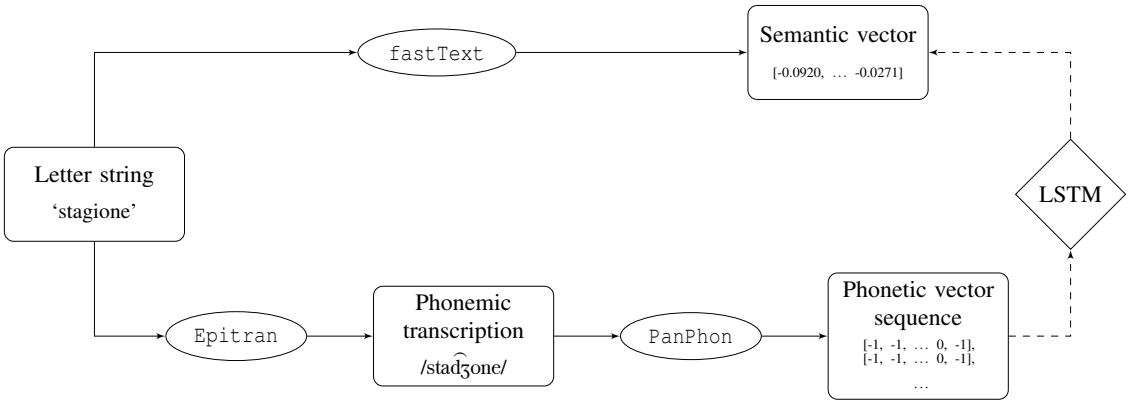


Figure 1: Schematic representation of the experimental pipeline

dense layers matched the dimensionality of the input and output vectors, respectively, whereas the LSTM layer was configured with 200 units, a dropout of 0.2 and a recurrent dropout of 0.2. Cosine similarity was used as both objective function and metric, and the Adam optimization method was employed for training (Kingma & Ba, 2014), with learning rate set to 0.01. We adopted the *tanh* activation function for the output layer, since its codomain corresponds to the range (-1, 1), in which the semantic vectors are defined. The hyperparameters were set without tuning.

### Experimental conditions

The experimental conditions were characterized by different combinations of training and test sets. In the cross-family condition, the model was trained for one epoch on the Arabic, Hungarian, Indonesian, Thai, Vietnamese, Turkish, and Tamil datasets, and then tested in Italian. All the languages employed in the cross-family model belonged to different language families, as shown in Table 1[2]. In the mono-family condition, the model was trained for one epoch on the Bengali, Hindi, Polish, Ukrainian, Dutch, French, and Spanish datasets, and tested in Italian. All the languages in this condition were Indoeuropean, including the test set. In order to make meaningful and unbiased comparisons between the performances of the cross- and the mono-family models, we randomly selected a subset of the original data for the Indoeuropean condition, matching the size of the cross-family training set. In the monolingual condition, which defined the upper bound of the network's performance, the LSTM was trained and tested on different subsets of the Italian dataset, with a train-test split ratio of 0.2. In order to partially compensate for the different size of the training set (roughly one fourth of the multilingual sample), the monolingual model was trained for four epochs. To define a baseline for the evaluation of the models' performances, we trained a randomized equivalent of each experimental model described above by randomly shuffling the output vectors of the relevant condition. All the words whose orthographic form was present in the training set were removed from the corresponding test set, in order to prevent the cross-linguistic consistency of proper names and lexical borrowings to inflate our results [3].

## Results

Table 2 reports the results of the models paired with their random counterparts. The first column of the table specifies the experimental condition; the second and third columns present the average cosine similarity between the target semantic vector and the experimental (*exp*) or the random (*r*) model's prediction for every item in the test set. We evaluated the statistical significance of our results through a set of paired samples *t*-tests between the element-wise cosine similarity of the target semantic vectors with the vectors generated by the two alternative models. The last two columns of the table present the *t* statistic and the associated *p*-value for each of the contrasts evaluated by the test. All the experimental models outperformed their respective randomized baselines; as expected, we found a consistent negative association between the linguistic distance of the languages in the training and in the test set and the results of the models: the monolingual performance was stronger than the one achieved by the cross-family model, with an ample difference of 0.1913 in the metric. Moreover, the mono-family model scored between the results achieved by the monolingual and the cross-family models.

The results of the experimental models may reflect different levels of correspondence between phonetic form and meaning. In the monolingual condition, it is not possible to discern the effects of compositionality-induced regularities from proper phonosemantic systematicity; in the mono-family condition, another possible source of non-arbitrariness which is not phonosemantic in nature is the etymological relatedness between words in typologically close languages. Conversely, the features extracted by the cross-family network can be considered phonosemantic in a narrow sense, being etymology- and compositionality-agnostic. The above-chance performance of the cross-family network is in line with our

[2]Following the Omniglot genealogical classification of languages at `https://omniglot.com/writing/langfam.htm`

[3]We thank Reviewer #1 for bringing this issue to our attention.

| Model | $N_{train}$ | $N_{test}$ | $Cos_{exp}$ | $Cos_r$ | $t$ | $p$ |
|---|---|---|---|---|---|---|
| Monolingual | 691259 | 172815 | 0.5105 | 0.3411 | 534.15 | $\ll 0.001$ |
| Mono-family | 2742222 | 331475 | 0.3882 | 0.3383 | 242.59 | $\ll 0.001$ |
| Cross-family | 2742222 | 558791 | 0.3192 | 0.2986 | 126.11 | $\ll 0.001$ |

Table 2: Test results by experimental condition

predictions, and consistent with the hypothesis that a certain degree of cross-linguistic correspondence between phonetic and semantic representations is already encoded in language; moreover, it shows that, with sufficient training, this correspondence can be efficiently captured by a computational system, and not only for a subset of (relatively) culture-independent concepts (Blasi et al., 2016; Wichmann et al., 2010), but at a lexicon-wide level.

## Follow-up analyses

In light of the results we reported, a natural question that arises is whether phonosemantic information is uniformly distributed in the lexicon, or some linguistic subspaces tend to incorporate stronger links with their phonetic realization. Our aim was to inspect the nature of these linguistic subspaces through a theory-driven quantitative analysis. In order to assess the potential influence of a set of linguistic factors on lexical phonosymbolism, we needed a proper metric to formalize the dependent variable. We chose to operationalize the degree of phonosemantic transparency of a word as the cosine similarity between the target semantic vector and the cross-family network's prediction for the items in the test set. The rationale under this choice was that since the network succeeded in the mapping from the phonetic to the semantic representational format, the words predicted with higher precision by the model would exhibit a higher rate of the phonosemantic features that the model managed to capture.

The following subsections illustrate the results of different regression analyses with various cognitively motivated predictors and the metric defined above as dependent variable. All the analyses described below include Twitter-based lexical frequency estimates (Gimenes & New, 2016) and word length as linear covariates.

### Semantic factors

Reilly et al. (2017) showed that the concreteness of a lexical item could be inferred with an above-chance accuracy by English-speaking participants in languages to which they had never been exposed. Aiming to dissect this composite perceptual variable in its inner constituents, we constructed five ordinary least squares regression models with the perceptual strength ratings for each of the five perceptual modalities provided by Vergallito et al. (2020) as predictors of lexical non-arbitrariness. The results are summarized in Table 3.

Although the variance explained by the models, expressed in terms of $R^2$, is consistently low, the results report a significant, positive effect of perceptual strength in the haptic and visual modalities, whereas the predictors based on the auditory, gustatory and olfactory ratings do not reach statistical significance[4]. Perceptual availability in the modalities receptive to plastic attributes (shape, position, orientation, depth) seems thus to be associated with higher phonosemantic transparency, even if lexical frequency and word length are controlled for.

| Sensory modality | $\hat{B}$ | $t$ | $p$ | $R^2$ |
|---|---|---|---|---|
| Auditory | -0.0019 | -1.049 | 0.295 | 0.008 |
| Gustatory | 0.0061 | 2.489 | 0.013 | 0.012 |
| Haptic | 0.0110 | 6.280 | $< 0.0001$*** | 0.041 |
| Olfactory | 0.0051 | 2.289 | 0.022 | 0.011 |
| Visual | 0.0092 | 3.442 | 0.001** | 0.017 |

Table 3: Results of the regression models with perceptual strength ratings as predictors of phonosemantic transparency. The asterisks indicate the statistical significance of the model after a Bonferroni correction is applied on the $\alpha$-level (* = 0.05/5; ** = 0.01/5; *** = 0.001/5). The sample is composed by the items that were present in the test set, the Italian perceptual norms, and the frequency estimates (N = 1092).

### Syntactic factors

Systematic cross-linguistic studies (Woodworth, 1991; Johansson & Jordan, 2013) have shown a consistent correspondence between spatial orientation and phonological realization for the grammatical class of demonstratives. In order to verify the generalizability of this idea to the superclass of function words, we trained a unigram tagger on a 100M subset of the Paisà corpus (Lyding et al., 2014), and derived the coarse-grained POS-tags for the items in our test set. We then collapsed the obtained tags into two superclasses as content (adjectives, nouns, verbs) and function words (conjunctions, determiners, prepositions, interjections, numerals, pronouns, articles, predeterminers). The items that did not fall unambiguously into the previous classes (adverbs, non-tagged words) were excluded from the analyses. We ran a linear regression to predict the previously defined index of lexical non-arbitrariness from the grammatical superclass of the word, with content words dummy-coded as 0 and function words as 1 (N = 54305). Although the amount of variance explained by the model was low ($R^2 = 0.005$), the positive value of the unstandardized regression coefficient ($\hat{B} = 0.0211$) and its high statistical significance ($t = 5.942$, $p < 0.001$) provide empirical support for the hypothesis that function words in general might be associated with a privileged link between sound and meaning.

### Pragmatic factors

We aimed to extend our analysis of the linguistic variables that affect lexical phonosymbolism to pragmatic factors; with the limitations of a word-level study, we directed our inquiry towards interjections. Interjections should be regarded as

[4]The values of the overall significance of the models are not considered to be of theoretical interest, since they include the effects of the linear covariates.

universal pragmatic markers (Norrick, 2009). They express spontaneous feelings or reactions (Bloomfield, 1984) and can be closely related to their natural manifestation (Wharton, 2003); hence, it would not be surprising to find a more transparent link between their phonoarticulatory expression and their meaning.

In order to investigate the hypothesis of a favoured phonosemantic mapping in interjections, we ran a linear regression to test whether a word being an interjection or not could predict its phonosemantic transparency. We collapsed the unigram tags for all the word classes that were not interjections, and binarily coded interjections as 1, and non-interjections as 0 (N = 54305). Again, the amount of variance explained by the model was low ($R^2 = 0.006$), but the positive unstandardized coefficient ($\hat{B} = 0.0711$) and the regressor's high statistical significance ($t = 6.702$, $p < 0.001$) provide experimental evidence for a privileged link between the pragmatic valence of a word and its degree of phonosymbolism. Moreover, this result is consistent with various anecdotal findings documented in the literature. For instance, Winter et al. (2017) reported that interjections are judged as the most non-arbitrary Parts of Speech by English speakers. Additional converging evidence comes from a cross-linguistic approach to a single pragmatic marker: Dingemanse et al. (2013) showed that the interjection "Huh?" is a universal, found in roughly the same form and function in spoken languages across the globe.

### Developmental factors

Non-arbitrariness has been integrated into different theories of language acquisition (Asano et al., 2015; Imai et al., 2008; Massaro & Perlman, 2017) in order to alleviate Quine's logically insurmountable problem of linking the phonological form of a novel word with its meaning (Quine, 1960), and the speech segmentation (or word discovery) problem, i.e. the initial difficulty in the localization of word boundaries in a continuous speech stream without the knowledge of any word. Massaro & Perlman (2017) demonstrated empirically that phonosemantic transparency is more prevalent in the lexicon at early acquisition stages, later diminishing with increasing age and vocabulary. We aimed to inspect whether this developmental tendency was reflected by a broad association between phonosemantic transparency and the age of acquisition of a word. We conducted a regression analysis relying on the Italian age of acquisition norms released by Montefinese et al. (2019). The regressor of interest reached high statistical significance ($t$ = -8.236, $p < 0.001$), although the model explained a low portion of the variance ($R^2 = 0.042$, with N = 1946). The negative regression coefficient ($\hat{B}$ = -0.0067) confirmed our expectation that words learned in earlier stages of the acquisition of the lexicon tend to be associated with higher phonosemantic transparency, in line with previous behavioral results. Phonosymbolic links could then help children learn semantic concepts, and discover structures across spoken and contextual input.

## Discussion

Most of our findings align with the psycholinguistic literature on non-arbitrariness: a negative relationship with age of acquisition has been reported by Massaro & Perlman (2017) and Perry et al. (2015), among others. Furthermore, Winter et al. (2017) have shown that words with meanings related to the senses display a stronger link with their phonetic realization than words with abstract meanings, and that interjections achieve the highest ratings of form-to-meaning transparency. Nonetheless, a few major differences between our results and the main trends in the aforementioned psycholinguistic studies are worthy of mention. Among sensory modalities, Winter et al. (2017) found that auditory and tactile words were considered to be less arbitrary than those related to the other senses; Perry et al. (2015) and Winter et al. (2017) showed that adjectives were rated as more iconic than function words. To address this apparent contradiction, we wish to highlight a profound difference between our operationalization of the construct and the one employed in these studies: while for the latter non-arbitrariness was assessed through *explicit* ratings, where participants were asked to evaluate how iconic a word sounded in a specific language, we employed an *implicit* measure of how well the phonetic representation of a word could enable a cross-linguistic network to infer its meaning. We deem that our novel measure of language-independent non-arbitrariness could complement traditional explicit measurements in the investigation on how languages relate sound to meaning.

The conflict between the phonosemantic transparency of function words on the one hand and of perceptually available terms on the other might raise the problem of how these two tendencies can coexist in their apparent contradiction. The two trends are not directly in contrast within our study, since the perceptual norms on which we performed our semantic analyses only contained content words; hence, our results could simply reflect the effect of a concreteness gradient limited to content words. Nonetheless, we speculate that these two opposite effects might be the result of two distinct tendencies, namely effectiveness and efficiency. For content words, it is *efficient* to establish a link between sounds and referents if the latter have salient physical attributes that can be related to linguistic sounds. For function words cementing such links is more demanding, but at the same time their presence is more *effective* in the context of language learning. Indeed, for most function words there is no possibility of learning by ostension, and a higher transparency might be more beneficial in the acquisition of the lexicon. Moreover, the results of the analyses on semantic and syntactic factors might reflect different facets of non-arbitrariness. In the former case, the detected effects might reflect instances of iconicity, a form of non-arbitrariness in which aspects of the form and meaning of words are related by means of perceptuomotor analogies; in the latter case, the higher phonosemantic transparency of function words might be driven by systematicity, a different form of non-arbitrariness prompted by statistical regularities

between sound and usage patterns of word classes (Dingemanse et al., 2015). Dingemanse et al. (2015) suggest that the phonological cues that help in discerning between word classes might be language-specific, and characterized by ample cross-linguistic differences. We believe that our results concerning function words provide empirical evidence against this view; the favoured form-to-meaning mapping in function words cannot be considered as iconic in a narrow sense – with the exception of spatial demonstratives –, and yet, the relationships between their phonological profiles and their relative position in the embedding space can be transferred across different language families.

The findings presented in the previous subsections suggest that phonosemantic information is not uniformly distributed in the lexicon: the consistency of the mapping between sound and meaning seems to be influenced by semantic, syntactic, pragmatic, and developmental factors. We remark that the list of variables examined in the present study is not exhaustive, and we leave for future research the assessment of other linguistic factors entangled with lexical phonosymbolism.

## Conclusion

In the present study, we showed that an LSTM model trained in a cross-linguistic setup can identify a possibly universal sound-symbolic substrate underlying diverse language families, and yield language-independent generalizations in the mapping from sound to meaning. Our results substantiate the claim that linguistic phonosymbolism, being entangled with language at different levels of analysis, should be regarded as a widespread linguistic phenomenon. While in the present work we dissected the semantic space to find privileged phonosemantic regions, we leave for future research an assessment of the precise contribution of the phonetic features that shape the iconic correspondences permeating the lexicon.

## References


Abramova, E., & Fernández, R. (2016, June). Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 343–352). San Diego, California: Association for Computational Linguistics. doi: 10.18653/v1/N16-1038

Abramova, E., Fernández, R., & Sangati, F. (2013). Automatic labeling of phonesthemic senses..

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *cortex*, *63*, 196–205.

Berlin, B. (1995). Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (p. 76–93). Cambridge University Press. doi: 10.1017/CBO9780511751806.006

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823. doi: 10.1073/pnas.1605782113

Bloomfield, L. (1984). *Language*. University of Chicago Press. (Google-Books-ID: 87BCDVsmFE4C)

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, *126*(2), 165 - 172. doi: https://doi.org/10.1016/j.cognition.2012.09.007

Chollet, F., et al. (2015). *Keras.* `https://keras.io`.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149-2169. doi: 10.1111/cogs.12453

Dingemanse, M., Blasi, D., Lupyan, G., Christiansen, M., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*, 603-615. doi: 10.1016/j.tics.2015.07.013

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE*, *8*(11). doi: 10.1371/journal.pone.0078273

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972. doi: 10.3758/s13428-015-0621-0

Gutiérrez, E. D., Levy, R., & Bergen, B. (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)* (pp. 2379–2388). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16-1225

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65.

Jakobson, R., & Waugh, L. R. (2011). *The sound shape of language*. Walter de Gruyter.

Johanssohn, N., Anikin, A., & Aseyev, N. (2020). Color sound symbolism in natural languages. *Language and Cognition*, *12*(1), 56–83. doi: 10.1017/langcog.2019.35

Johansson, N., & Jordan, Z. (2013). Motivations for sound symbolism in spatial deixis: A typological study of 101 languages. *Public Journal of Semiotics*, *5*, 3-20.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the*

*2018 conference on empirical methods in natural language processing*.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.*

Köhler, W. (1929). *Gestalt psychology*. Liveright.

Koriat, A., & Levy, I. (1977). The symbolic implications of vowels and of their orthographic representations in two natural languages. *Journal of Psycholinguistic Research*, *6*(2), 93–103. doi: 10.1007/bf01074374

Lockwood, G., & Tuomainen, J. (2015). Ideophones in japanese modulate the p2 and late positive complex responses. *Frontiers in psychology*, *6*, 933.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, *7*(2), 167–193.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., ... Pirrelli, V. (2014). The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th web as corpus workshop (WaC-9)* (pp. 36–43). Gothenburg, Sweden: Association for Computational Linguistics. doi: 10.3115/v1/W14-0406

Massaro, D. W., & Perlman, M. (2017). Quantifying iconicity's contribution during language acquisition: Implications for vocabulary learning. *Frontiers in Communication*, *2*, 4. doi: 10.3389/fcomm.2017.00004

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316-322. doi: 10.1111/j.1467-7687.2006.00495.x

Monaghan, P., Shillcock, R., Christiansen, M., & Kirby, S. (2014). How arbitrary is language? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*. doi: 10.1098/rstb.2013.0299

Montefinese, M., Vinson, D., Vigliocco, G., & Ambrosini, E. (2019). Italian age of acquisition norms for a large set of words (itaoa). *Frontiers in Psychology*, *10*, 278. doi: 10.3389/fpsyg.2019.00278

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (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* (pp. 3475–3484). Osaka, Japan: The COLING 2016 Organizing Committee.

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of Pragmatics*, *41*(5), 866 - 891. (Pragmatic Markers) doi: https://doi.org/10.1016/j.pragma.2008.08.005

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, *28*(1), 225-252. doi: 10.1146/annurev.anthro.28.1.225

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, *214*(3), 373–380.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PloS one*, *10*(9).

Quine, W. (1960). *Word and object: An inquiry into the linguistic mechanisms of objective reference*. Oxford, England: John Wiley.

Rabaglia, C. D., Maglio, S. J., Krehm, M., Seok, J. H., & Trope, Y. (2016). The sound of distance. *Cognition*, *152*, 141–149.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Reilly, J., Hung, J., & Westbury, C. (2017). Non-arbitrariness in mapping word form to meaning: Cross-linguistic formal markers of word concreteness. *Cognitive Science*, *41*(4), 1071-1089. doi: 10.1111/cogs.12361

Sagi, E., & Otis, K. (2008). Semantic glimmers: Phonaesthemes facilitate access to sentence meaning..

Sathian, K., & Ramachandran, V. S. (2019). *Multisensory perception: from laboratory to clinic*. Elsevier.

Saussure, F. d. (1964). Course of general linguistics (cours de linguistique générale, 1959). second impression. ed. by charles bally and albert sechehaye. *Trans. Wade Baskin. London: Peter Owen*.

Shillcock, R., Kirby, S., & McDonald, S. (2001). Filled pauses and their status in the mental lexicon.

Slavova, V., et al. (2019). Towards emotion recognition in texts–a sound-symbolic experiment. *International Journal of Cognitive Research in Science, Engineering and Education*, *7*(2), 41–51.

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*, 259-278.

Vainio, L., Schulman, M., Tiippana, K., & Vainio, M. (2013). Effect of syllable articulation on precision and power grip performance. *PloS one*, *8*(1), e53061.

Vergallito, A., Petilli, M., & Marelli, M. (2020). Perceptual modality norms for 1,121 italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*, *52*. doi: 10.3758/s13428-019-01337-8

Wharton, T. (2003). Interjections, language, and the 'showing/saying' continuum. *Pragmatics & Cognition*, *11*, 39-91. doi: 10.1075/pc.11.1.04wha

Wichmann, S., Holman, E., & Brown, C. (2010). Sound symbolism in basic vocabulary. *Entropy*, *12*. doi: 10.3390/e12040844

Winter, B., Perlman, M., Perry, L., & Lupyan, G. (2017). Which words are most iconic? iconicity in english sensory words. *Interaction Studies*, *18*. doi: 10.1075/is.18.3.07win

Woodworth, N. L. (1991). Sound symbolism in proximal and distal forms. *Linguistics*, *29*(2), 273 - 300. doi: https://doi.org/10.1515/ling.1991.29.2.273