




Cognitive Science 46 (2022) e13147

© 2022 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.13147

# A Cross-Modal and Cross-lingual Study of Iconicity in Language: Insights From Deep Learning

Andrea Gregor de Varda,<sup>a</sup>  Carlo Strapparava<sup>b</sup>

<sup>a</sup>*Department of Psychology, University of Milano – Bicocca*

<sup>b</sup>*FBK – Fondazione Bruno Kessler*

Received 25 November 2021; received in revised form 27 April 2022; accepted 28 April 2022

---

## Abstract

The present paper addresses the study of non-arbitrariness in language within a deep learning framework. We present a set of experiments aimed at assessing the pervasiveness of different forms of non-arbitrary phonological patterns across a set of typologically distant languages. Different sequence-processing neural networks are trained in a set of languages to associate the phonetic vectorization of a set of words to their sensory (Experiment 1), semantic (Experiment 2), and word-class representations (Experiment 3). The models are then tested, without further training, in a set of novel instances in a language belonging to a different language family, and their performance is compared with a randomized baseline. We show that the three cross-domain mappings can be successfully transferred across languages and language families, suggesting that the phonological structure of the lexicon is pervaded with language-invariant cues about the words' meaning and their syntactic classes.

*Keywords:* Non-arbitrariness; Phonosymbolism; Iconicity; Cross-lingualism; Language and vision; Deep learning

---

## 1. Introduction

A pivotal property of human languages is their ability to refer to entities and events that populate the physical world by means of signs. In oral languages, these signs consist of ordered sequences of sounds; the links between these phonological patterns and the world are determined by both the phonemes that are uttered and their relative position within a

---

Correspondence should be sent to Andrea Gregor de Varda, Department of Psychology, University of Milano – Bicocca, Piazza dell'Ateneo Nuovo 1, Milano, MI 20126, Italy. E-mail: a.devarda@campus.unimib.it

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

word. The nature of the relationships that tie the speech sounds composing a word and that word's meaning has kindled the interest of philosophers since ancient times (see Magnus, 2013, for a historical overview); nonetheless, the deep-rooted fascination for this puzzling question was waned by the empiricist criticism put forward by Locke (1847) and definitively annihilated by the structuralist axiom of the arbitrariness of the sign (Saussure, 1964). According to the Saussurean perspective on meaning, words should be conceived as arbitrary labels, forced onto the semantic concept they refer to as a result of social processes of cultural transmission. This framework quickly conquered the theoretical panorama (see, for instance, Bloomfield, 1984; Hockett & Hockett, 1960; Levelt, Roelofs, & Meyer, 1999), rejecting a priori the possibility of any natural correspondence between the linguistic sounds and their denotation. However, as noted by Allott (2001), this perspective has restrained the study of the phonological properties of the lexicon beyond the reach of scientific explanation.

The concept of iconic referentiality dismisses the assumption of an arbitrary link between the words and their *denotatum*; it entails that linguistic sounds can bear meaningful associations with their referents, with these associations being mediated not only by the phonological regularities of a given language but also by the sounds' inherent qualities (i.e. their acoustic and articulatory features). Approaches that incorporate iconic principles into lexical semantics are gaining increasing popularity in cognitive research, but they are still conceived as an alternative to the standard view on vocabulary structure. Notably, motivated mappings with the phonological form seem to be rejected a priori only in the lexical domain. However, it is commonly recognized that non-arbitrary cross-domain connections account for a variety of linguistic phenomena at different levels of analysis beyond the study of the lexicon. From a syntactic standpoint, it is widely acknowledged that linguistic structures mirror various facets of the structure of experience (Croft, 2002; Haiman, 1985; Levinson, Stephen, & Levinson, 2000). The parallelism between linguistic and temporal sequences has been proposed as an example of this correspondence (Bybee, 1985; Perniss, Thompson, & Vigliocco, 2010): in the sentence "I will eat, shower, and read a book" the hearer will typically infer that the speaker intends to perform the three actions in the order in which they were uttered. Nonetheless, there is no external cue besides the sequential arranging of the verb phrases that support this assumption. Proposals that incorporate iconic principles in linguistic analysis have also been outlined in the domain of morphology, with the observation that for degree adjectives (e.g., *big*, *bigger*, *biggest*) the highest degree of quality is iconically represented by the word with the greatest number of phonemes in its inflection (Wescott, 1971). The claim that a given domain can be structured without any accountable principle is inherently sterile. For this reason, the first research efforts that challenged this view were welcomed with a high resonance in the scientific community.

### 1.1. Phonovisual iconicity

In the late 1920s, the cognitive sciences drew attention to some anecdotal cases that challenged the structuralist principle of the arbitrariness of the sign. Two prominent studies disclosed a non-trivial link between the participants' guesses about a figure's name and some of its visual properties, namely its shape (Köhler, 1929) or its size (Sapir, 1929). In Sapir's study

(1929), participants engaged in a name matching task; they were presented with the images of two tables of different sizes and instructed to pair them with the pseudowords “mil” and “mal.” Intriguingly, the latter phonetic sequence was coupled four times more often with the larger object, showing that the participants’ intuitions were biased by the nature of the vocalic phone. In the same year, Köhler showed that the phonological profiles of two non-words affected their association with two novel shapes: participants tended to label “maluma” a rounded shape and “takete” a spiky one. This latter experiment on shape phonosymbolism had a wide impact on experimental psychology and linguistics: Several studies managed to replicate Köhler’s results, corroborating the psychological reality of the so-called “maluma–takete” effect (Köhler, 1947; Werner, 1948, 2011) or “bouba-kiki” effect, referring to the pseudowords employed by Ramachandran & Hubbard (2001). These research efforts set the stage for a number of experiments that repeatedly reported the same phonovisual correspondences at different developmental stages (Maurer, Pathman, & Mondloch, 2006; Ozturk, Krehm, & Vouloumanos, 2013; Pejovic & Molnar, 2017) and in various linguistic, geographical, and cultural contexts (Bremner et al., 2013; Ramachandran and Hubbard, 2001; Chen, Huang, Woods, & Spence, 2016; Shinohara & Kawahara, 2010; Ćwiek et al., 2021). The results of the studies on shape and magnitude symbolism were complemented by other findings that related to different properties of the visual modality, such as color (Johanssohn, Anikin, & Aseyev, 2020) and lightness (Hirata, Ukita, & Kita, 2011), to their respective phonetic signs.

Vision is not the only sense by which we experience the world, and several studies have searched for a phonosensory bias in different perceptual modalities. Iconic sensory analogies were then documented in various senses, such as touch (Fryer et al., 2014; Graven & Desebrock, 2018), smell (Atkinson, Speed, Wnuk, & Majid, 2021), kinesthesia (Fontana, 2013), and taste (Gallace, Boschini, & Spence, 2011). Iconic words that make reference to the auditory modality are particularly relevant in linguistic and cognitive research, since their phonosymbolic mapping takes place *within* a modality, relating verbal and non-verbal sounds. They receive the highest explicit iconicity ratings (Winter et al., 2017), and participants are able to associate them with their meaning with the highest accuracy among all the other sensory modalities (Dingemanse, Reinisch, Schuerman, Tufvesson, & Mitterer, 2016). In the ideophonic lexicon – i.e the portion of the vocabulary that includes marked words depicting sensory imagery (Dingemanse, 2012) – auditory terms are the most prominent class. They occupy the highest rank in the cross-linguistic implicational hierarchies developed by Blasi, Dingemanse, Lupyan, Christiansen, and Monaghan (2015) and revised by McLean (2021), meaning that if a language does not develop auditory ideophones, it will not produce ideophones related to the other senses. Nonetheless, we chose to focus on vision since we were interested in an analogical iconic mapping that involved a cross-modal link. With the exception of the auditory modality, the phonovisual biases hold a privileged role among the other senses, both in terms of the research interest they elicited and the consistency of the findings. Indeed, the cross-modal correspondences in the olfactory-gustative modality do not seem to be coherent across cultures (Bremner et al., 2013), and the iconic biases in the haptic domain might be mediated by visual imagery (Fryer et al., 2014) and auditory experiences (Winter et al., 2017). In the light of this asymmetry, we approached the multifaceted subject of

perceptual iconicity by analyzing the link between phonological profiles and visual features (Experiment 1).

### 1.2. *Phonosemantic iconicity*

Visual representations do not exhaust the whole semantic spectrum. Several words are grounded in other perceptual modalities, and abstract concepts lack a precise relationship with sensorimotor features in general (Borghi et al., 2017; Crutch & Warrington, 2005; Lupyan & Winter, 2018; Paivio, 2010). Not only are sounds associated with other sensory properties, but they are also more generally associated with lexical meanings. For example, studies have shown that participants are able to couple with an above-chance accuracy visually presented characters (Koriat & Levy, 1977) and auditorily presented words (Berlin, 1995) of a foreign language with their meaning. 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, Hung, & Westbury, 2017). Taken together, these findings suggest that the semantic information encoded in a word's phonological profile may include other features that are not exclusively visual. Aiming to extend the scope of our study beyond the domain of perception, we devised a second experiment where we inspected the link between sound and language-based meaning representations (Experiment 2).

### 1.3. *Systematicity*

Dingemanse et al. (2015) drew an important distinction between two patterns of non-arbitrariness in vocabulary structure, namely iconicity and systematicity. The former term reflects the idea that phonemes can convey meaning *per se*, that is, not only through contrastive relations with other sounds but also through their intrinsic sound qualities; in iconic words, aspects of form and meaning are related by means of perceptuomotor analogies. The latter constitutes a different form of non-arbitrariness prompted by statistical regularities between sound and usage patterns of word classes. Despite its pervasiveness, systematicity has received relatively little attention in linguistics and cognitive science. Systematicity does not concern a direct relationship between phonetic patterns and referential semantic properties; instead, it regards the phonetic regularities that are instantiated within a word class. Class-level phonetic cues have been found in a broad range of languages, with evidence coming from both typological (Smith, 2011) and corpus studies (Monaghan, Christiansen, & Chater, 2007). Dingemanse et al. (2015) suggested that the phonetic cues that help in discerning between word classes might be language-specific, featuring ample cross-linguistic differences. The results from our study (Experiment 3) challenge this assumption, providing empirical evidence that the relationship between phonological profiles and word classes is characterized by significant cross-linguistic consistency and can be transferred across different language families. Within the computational framework, the analysis of lexical non-arbitrariness has largely focused on iconicity (Abramova & Fernández, 2016; Abramova, Fernández, & Sangati, 2013; Blasi, Hammarström, Wichmann, Stadler, & Christiansen, 2016; Johansson, Anikin, & Aseyev, 2020; Shillcock, Kirby, & McDonald, 2001; Wichmann,

Holman, & Brown, 2010; although see Gutiérrez, Levy, & Bergen, 2016; Monaghan et al., 2007; Tamariz, 2008). We believe that the eventual cross-linguistic consistency of the systematic cues that help in distinguishing between word classes deserves to be addressed at a large scale, and our third experiment aims to fill this research gap.

#### 1.4. *Relevance*

In recent years, non-arbitrariness has gone from being merely peripheral to the interests of the cognitive science community to being integrated into broader theories of language evolution (Cabrera, 2012; Dingemans et al., 2013; Ramachandran and Hubbard, 2001), processing (Lockwood & Tuomainen, 2015) and acquisition (Asano et al., 2015; Imai, Kita, Nagumo, & Okada, 2008; Murgiano, Motamedi, & Vigliocco, 2021). A naturally biased relationship between phonetics and semantics restrains the problem space of the evolution of language, posing constraints on the emergence of the vocabulary. Furthermore, a systematic link between a linguistic sound and its referent might strengthen the mnemonic traces in the process of language acquisition (Sathian & Ramachandran, 2019). The effects of phonosemantic correspondences are not encased within language but have been shown to spread to different cognitive faculties, such as categorization (Lupyan & Casasanto, 2015), memory (Ramachandran and Hubbard, 2001), and emotion recognition (Slavova, 2019); moreover, they exert an influence on actional processes such as phonatory behavior (Parise & Pavani, 2011), spatial navigation (Krehm, Maglio, Rabaglia, Seok, & Trope, 2016), and hand grip (Schulman, Vainio, Tiippana, & Vainio, 2013). In the light of their effects within the human cognitive system, the phonosemantic biases are not likely to be limited to a few circumscribed phonetic or semantic clusters, but may instead pervade the lexicon beyond the often reported anecdotal instances.

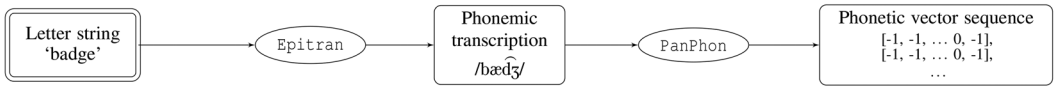
#### 1.5. *Aims*

In the present study, we tackle the following questions:

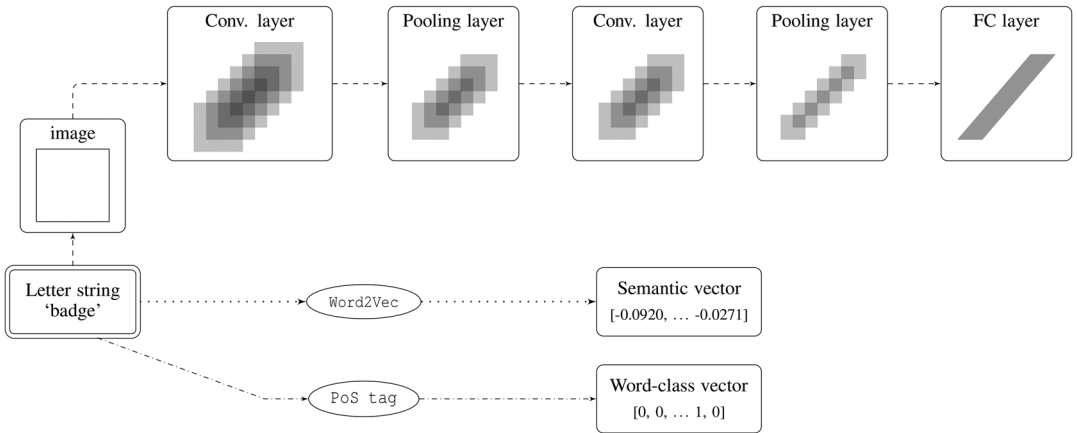
- Is there a relationship between the phonological realization of a word and the visual representation of its referent?
- Is this sound-to-meaning link extended beyond visual semantics?
- Are word classes organized into consistent phonological clusters?

In trying to answer these inquiries, we adhered to three core methodological choices. First, we relied on large-scale data-driven procedures, with the intention of assessing the pervasiveness of non-arbitrariness in a representative linguistic sample, without including human biases in the item selection. Second, we implemented our experiments in a cross-linguistic setting. We deem that cross-linguistic diversity is a pivotal testbed for testing the hypothesis of a universal sound-symbolic substrate underlying all languages, as opposed to language-specific idiosyncratic systematicity. Third, we configured our experiments as zero-shot cross-lingual transfer learning tests, where we trained different long short-term memory (LSTM)-based recurrent neural networks in associating phonetic vector sequences with visual (Experiment 1), semantic (Experiment 2), and word-class (Experiment 3) representations (see Fig. 1).

① INPUT REPRESENTATIONS



② OUTPUT REPRESENTATIONS



③ CROSS-DOMAIN MAPPING

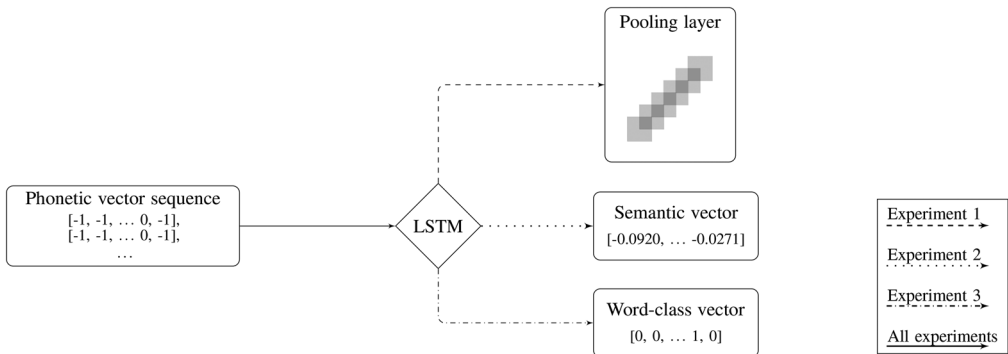


Fig. 1. Graphical summary of the experimental pipeline. The sequence in ① represents the stages of the graphemic-to-phonetic conversion, which is common to our three experiments. The flowchart in ② depicts the pre-processing stages of the images (Experiment 1, upper part of the diagram), the semantic vectorization of the words through Word2Vec (Experiment 2, middle part of the diagram), and the encoding of the word classes (Experiment 3, lower part of the diagram). For typographical reasons, only five layers of the VGG16 network are graphically depicted. The schema in ③ represents the cross-domain mappings from the phonetic input derived in ① to the output representations obtained in ②. Again, the representations of Experiment 1 occupy the higher position in the diagram, the ones of Experiment 2 are in the middle, and the ones of Experiment 3 are depicted at the bottom (see the legend on the right).

We followed the rationale that if the semantic and syntactic traces contained in the phonetic realization of the lexicon are consistent across languages – hence being a language universal in a broad sense –, then it should be possible to learn a cross-domain mapping in a set of unrelated languages and transfer it to a novel, typologically distant language without further training.

## 2. Experiment 1

In our first experiment, we explored the idea that vocabulary might be entangled with sensory experience. More precisely, we tested the hypothesis that the phonological properties of the lexicon might be related to the visual world by means of cross-modal correspondences, and that these correspondences might be consistent across languages. In order to uncover hidden links between these two domains – a task that arguably requires the use of complex transformations –, we trained an LSTM network to associate phonetic vector sequences with visual vectors denoting their referents. The latter were obtained through a forward pass of an image through a pre-trained hierarchical convolutional neural network (henceforth HCNN; see Section 2.4). The experimental pipeline is summarized in Fig. 1 (dashed line).

### 2.1. Dataset

We performed our first experiment on the THINGS dataset (Hebart et al., 2019), a resource that comprises 26,107 high-quality naturalistic images depicting a set of 1,854 diverse object concepts. Each item of the dataset was composed of an image paired with a label; these two components were pre-processed independently, as described in Subsections 2.3 and 2.4. In order to restrict the effects of the morphological noise in the labels, we removed from the dataset all the compound words (305 concrete words, corresponding to 3,839 images).

### 2.2. Translation

Each image label in the resulting dataset was translated into five languages belonging to five language families (see Table 1). We are aware that the choice of translating the labels is not free of concerns: the translation process does not always return the exact same concept in a different language, but rather the concept that overlaps to the highest degree with the original one. However, translating the labels licenses meaningful comparisons across languages – at least, more than employing different language-specific datasets –, and allows us to align the lexical items across languages, a crucial aspect to take into account when devising disjoint experimental conditions (see Section 2.6). In order to maximize the cross-lingual coverage of our dataset, while at the same time maintaining a high-quality translation, we first searched for lexical matches through word2word (Choe, Park, & Kim, 2020), a collection of bilingual lexicons constructed from the publicly available OpenSubtitles2018 dataset (Lison, Tiedemann, & Kouylekov, 2018); then, for the missing items, we employed the ground-truth bilingual dictionaries based on fastText, released by Facebook Research (Conneau, Lample, Ranzato, Denoyer, & Jégou, 2017). We removed from the analyses the words for which





a translation was missing in one or more languages – in other words, we considered only the set intersection of the translated items. The resulting dataset consisted of 16,820 images, depicting a set of 1,161 concrete words. The percentage of translations obtained with each translation tool for all the languages considered in the study is reported in Table 1, along with the percentage of missing items.

### 2.3. *Phonetic representations*

We derived the phonetic vector sequences corresponding to each image's label through the Epitran-PanPhon pipeline. In the first step of the procedure, the orthographic text of the label was transliterated into the International Phonetic Alphabet (IPA) with Epitran, a Python package for phonemic transcription. Then, the output was translated into a sequence of feature vectors with PanPhon, a library that converts IPA segments into subsegmental articulatory features (Mortensen et al., 2016). In line with Jakobson and Waugh (2011), who state 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), we chose not to directly hot-encode the IPA strings. In our opinion, the information-rich representational format offered by a phonetic feature decomposition is desirable since it uncovers the internal asymmetries that make different phones more or less related to each other (see Blasi et al., 2016; Joo, 2020, for similar considerations). With a hot-encoding over the IPA vocabulary all the phones would correspond to discrete categories, while two phones might differ by a single feature (e.g., [p] and [b], which are only distinguished by the feature [+/- voiced]), or more than 10 (e.g., [t] and [u], which exhibit 13 different subsegmental features).

The words in the input could be composed of a variable number of phones, which would result in vector sequences of different lengths. To make the tensor shapes comparable, all the input sequences were zero-padded, with a maximum length of 15. Thus, vector sequences derived from words with less than 15 phones were extended with zeroes, which would be subsequently hidden in the masking layer of the LSTM network (see Section 2.5). On the other hand, vector sequences corresponding to words with more than 15 phones were truncated, and the phonetic vectors corresponding to the following phones were discarded. Note that the items with 15 phones or more were less than 0.03% of the total, so the number of truncated words was negligible.

### 2.4. *Visual representations*

To transform the raw RGB images in the input into cognitively inspired visual representations, we relied on VGG16, an HCNN for large-scale image recognition (Simonyan & Zisserman, 2015). HCNNS exploit the hierarchical nature of the visual data to assemble representations of increasing complexity using small and simple patterns repeated across the images in input. They are biologically inspired models (LeCun & Bengio, 1995; LeCun et al., 2015) that have been developed in the field of computer vision with the purpose of classifying images, predicting a label from the pixel-wise RGB codes in input (Krizhevsky, Sutskever, & Hinton, 2012). HCNNS are usually composed of stacked convolutional and pooling layers, followed by standard fully connected (FC) layers (Simonyan & Zisserman, 2015). Convolu-

tional layers create feature maps that represent in a distributed tensor format the presence of features of various levels of abstraction in the input; these features are extracted through the application of learned filters to input images. Pooling layers then serve the purpose of lowering the resolution of the feature maps. Since the absolute position of a certain feature forming a motif might vary, coarse-graining each feature's position can create invariance to small shifts and distortions (LeCun et al., 2015). In deep models, shallow layers usually learn low-level visual features (e.g., lines, edges, color blobs) while layers that are deeply embedded in the network can extract high-level attributes (e.g., object parts, textures). The final layers then encode images as complex representations (e.g., object shapes) which are employed for the ultimate purpose of the network, that is, classification (Mahendran & Vedaldi, 2015).

The visual vectors included in this study consisted of the outputs of the fifth max-pooling layer of the pre-trained network, in response to a forward pass of each image in the dataset. After freezing all the model's weights by setting it in evaluation mode, we fed each image  $x$  in our stimulus set through the VGG16, in order to extract the resulting feature maps  $\varphi(x)$  of `block5_pool`; the resulting vectors were in turn flattened before being processed by the LSTM model. The weights of VGG16 were configured according to its pre-training on ImageNet (Deng et al., 2009). We employed the output of the VGG16 network as an approximation of a representational format proper to the human perceptual system. Indeed, HCNN-based representations can be successfully mapped onto neural responses to visual stimuli at different levels of processing within the ventral stream, even if the networks are not explicitly optimized to fit neural data (Yamins & DiCarlo, 2016). From a psychological perspective, these representations have been proposed to be cognitively plausible at least at the computational level of description (Marr, 1982), being able to predict human behavior and performance in several tasks (Günther, Petilli, Vergallito, & Marelli, 2020; Günther, Marelli, Tureski, & Petilli, 2021; Günther, Petilli, Vergallito, Ciapparelli, & Marelli, 2021).

### 2.5. Neural architecture

An LSTM was trained to associate the sequences of phonetic feature vectors in input into the visual vectors in output, with a many-to-one topological structure. The choice of the architecture was motivated by the nature of the input that LSTMs can process: while standard feed-forward neural networks can only treat single data points, LSTMs are endowed with feedback connections, that enable them to process inherently sequential data – in our case, the chains of phonetic vectors. The model was configured with Keras, a deep learning framework for Python (Chollet & others, 2015); it comprised a masking layer, followed by a single LSTM layer with 500 neural units, a dropout of 0.2, and a recurrent dropout of 0.2. The LSTM layer was connected to a dense layer with the number of units (25,088) matching the dimensionality of the target visual vector and equipped with rectification non-linearity (Rectified Linear Unit, *ReLU*). Cosine similarity was employed as both the objective function and metric, and the Adam optimization method was employed for training (Kingma & Ba, 2014), with the learning rate set to 0.01. All the hyperparameters described above were set without tuning. Random seeds were set for replicability purposes.

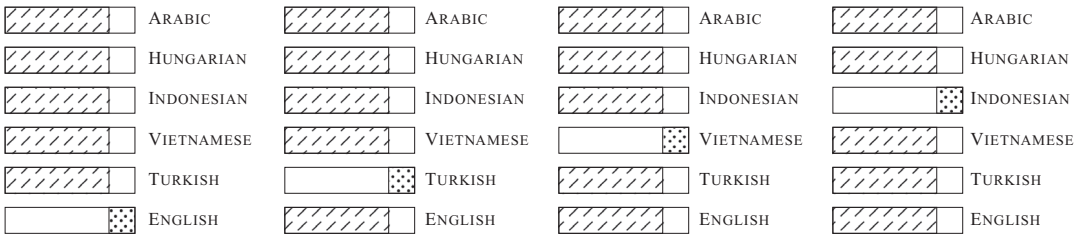


Fig. 2. Schematic summary of the train-test combinations in our experimental conditions. For typographical reasons, only four out of six conditions are reported. The rectangular shapes represent the totality of the word-image pairs available in each language. The training sets are marked with north-east lines, whereas the test sets are represented with the dotted patterns. As depicted in the figure, the training and test sets are completely disjoint. In the first block on the left, the training set is composed of a subset of the Arabic, Hungarian, Indonesian, Vietnamese, and Turkish data, whereas the test set is a different subset of English data.

## 2.6. Experimental conditions

We structured the experimental conditions following two main principles aimed at limiting the effects of the etymological relatedness of the items in the training and the test sets as much as possible. First, we prevented an image and a concept to occur in both sets, by randomly splitting the labels into two subsets, with a train-test split ratio of 0.8. Following this partition, the training set consisted of 929 concepts depicted by 13,397 images, whereas the test set was composed of 232 concepts represented by 3,423 images. Then, we devised our conditions so that the language on which the network was tested did not overlap with the set of languages on which the training was performed. In other words, in each condition the model was trained in the concatenation of the training sets of five languages  $\{L_i\}_{i=1,\dots,5}$  and tested in the test set of a sixth language  $L_6$  which was excluded from the training set, following a non-random sixfold cross-validation procedure (see Fig. 2). Each model was trained for one epoch in the five languages in the training set to map the phonetic vectorization of a word denoting an image to the convolution-based transformation of that image; then, it was tested on the same task in a novel language, belonging to a different language family. The experimental conditions were thus constructed so that the training and test sets were disjoint with respect to the concepts, the images representing them, and the languages to which the labels were translated. The experimental models' performances were assessed by comparing their results with the ones obtained by a parallel random model, which defined a baseline for quantifying the increase in performance due to the relevant multilingual signal. Concretely, the parallel model was trained on a dataset where the correspondence between input and output vectors was randomized by shuffling the visual vectors in output. All models were trained on 66,985 samples ( $13,397 \times 5$  languages) and tested on 3,423 items.

## 2.7. Results

The results of the six cross-lingual models and their random counterparts are reported in Table 2. The descriptive statistics in the first columns reveal that across all the experimental conditions, the cross-lingual models always outperformed their randomized baselines.

Table 2  
Results by experimental condition (Experiment 1)

Language	Cross-lingual Model			Randomized Baseline			Contrast		
	Cosine	<i>SD</i>	95% CI	Cosine	<i>SD</i>	95% CI	<i>t</i>	<i>p</i>	<i>d</i>
Arabic	0.2343	0.0411	[0.2329, 0.2357]	0.2243	0.0382	[0.2230, 0.2256]	10.42	≪ .0001	0.6600
Hungarian	0.2382	0.0410	[0.2368, 0.2396]	0.2278	0.0386	[0.2266, 0.2291]	10.78	≪ .0001	0.6321
Indonesian	0.2391	0.0394	[0.2378, 0.2404]	0.2243	0.0399	[0.2229, 0.2256]	15.45	≪ .0001	1.4385
Vietnamese	0.2320	0.0431	[0.2306, 0.2335]	0.2224	0.0384	[0.2211, 0.2237]	9.77	≪ .0001	0.4544
Turkish	0.2381	0.0404	[0.2367, 0.2394]	0.2257	0.0387	[0.2244, 0.2270]	12.99	≪ .0001	1.1956
English	0.2389	0.0418	[0.2375, 0.2403]	0.2228	0.0384	[0.2216, 0.2241]	16.60	≪ .0001	1.3220

*Note.* The first column of the table specifies the language of the fold on which the validation was performed, implying that the training had been carried out on all the languages but the one in the test set. The following six columns of the table present the mean, the standard deviation (*SD*), and the 95% confidence intervals (CI) of the cosine similarity between the target visual vector and the cross-lingual or the random model's prediction for every item in the test set. The last three columns indicate the statistics of the contrast between the two models' results (*t* statistic, *p*-value, and Cohen's *d*).

To test whether this pattern of results was associated with statistical significance, we contrasted the results of the random and the cross-lingual models through a set of paired samples *t*-tests between the element-wise cosine similarity of the target visual vector in output with the predictions of the two alternative models, for each experimental condition. The inferential tests confirmed the soundness of the descriptive results, with all the contrasts being statistically significant. Furthermore, across all the experimental conditions, the 95% confidence intervals (CI) of the cross-lingual models did not overlap with the corresponding CI of the parallel random models. This suggests that the sound-to-vision correspondences inherent in the phonological structure of the lexicon can be learned in any direction and generalized to all the languages included in our study. All the contrasts were associated with medium-to-large effect sizes, with the exception of the zero-shot transfer to the Vietnamese language.

## 2.8. Discussion

Our LSTM imaginative models, trained on multilingual data to induce a mapping between phonological profiles and sensory representations, showed the ability to learn cross-modal and cross-linguistic correspondences in the lexicon, suggesting that visual information is implicitly encoded in the phonological structure of linguistic data. Showing a link between *meaningful* speech sounds and visual representations, we complement the behavioral studies presented in the Introduction, which disclosed a powerful and consistent link between *meaningless* speech sounds and magnitude, color, and geometrical shape. Across our experimental conditions, the LSTM networks were able to engage in a generative process where their visual imagery reproduced real-world concrete representations better than what would be expected by chance. Our strict manipulation of the linguistic distance between the languages in the training and the test set allows us to rule out the effect of any etymological relatedness between the different languages' vocabularies.

### 3. Experiment 2

In our first experiment, we showed that the phonological structure of the concrete lexicon is entangled with a visual experience. Nonetheless, concrete and vision-related words constitute only a fraction of the whole vocabulary; additionally, even the words that refer to visually perceivable objects are not completely summarized by their visual attributes. For instance, the conceptual representation of the word *cloud* is not fully specified by a mental picture of a puffy white shape. We have auxiliary knowledge about their nature – for instance, the fact that they are mainly composed of water particles suspended in the atmosphere – we have acquired through language use and that is part and parcel of the word meaning. Aiming to extend our previous study beyond the perceptual domain, we trained an LSTM model to associate the phonetic with the corresponding language-based semantic representation of a word, encoded as a 100-dimensional word embedding. Word embeddings represent word meanings as high-dimensional vectors, usually extracted from large corpora of natural language data. These representations are rooted in the theoretical foundations of the distributional hypothesis, according to which the semantic similarity between words is a function of the similarity between the contexts in which they occur (Firth, 1957; Harris, 1954). Word embeddings do not consist of a coarse representation of the context in which the word occurs, but rather in an abstract structure that accumulates from encounters with lexical items and their context (Lenci, 2018). Different from visual vectors, they can be computed for every word in a corpus, and thus are suitable for extending our inquiry beyond the scope of visual semantics. The experimental pipeline of our second experiment is depicted in Fig. 1 (dotted line). The phonetic vectorization of the graphemic sequences in the input was identical in every aspect to our previous experiment. Hence, we redirect the reader to Section 2.3 for a detailed description of the procedure.

#### 3.1. Semantic representations

The semantic representations employed in this experiment consisted of pre-trained word embeddings generated with *word2vec* (Chen, Mikolov, Corrado, & Dean, 2013) from the British National Corpus, and released by Rei and Briscoe (2014). The representation learning tool was based on the skip-gram model, which inputs a sequence of words into a log-linear classifier with a continuous projection layer, trained to predict words within a window size of five. Semantic vectors were available for 1.93M words.

#### 3.2. Translation

The words for which a semantic vector was available were translated following the same pipeline as in Experiment 1. In this case, the procedure resulted in a severe data loss (see Table 1); nonetheless, the original dataset was much larger with respect to the THINGS database, and although it was not possible to obtain a translation for the vast majority of the items, the set intersection of the translations comprised 24,612 words.

Table 3  
Results by experimental condition (Experiment 2)

Language	Cross-lingual Model			Randomized Baseline			Contrast		
	Cosine	<i>SD</i>	95% CI	Cosine	<i>SD</i>	95% CI	<i>t</i>	<i>p</i>	<i>d</i>
Arabic	0.5263	0.0687	[0.5244, 0.5282]	0.5245	0.0654	[0.5227, 0.5263]	3.6849	.0002	0.0525
Hungarian	0.5253	0.0683	[0.5234, 0.5272]	0.5241	0.0665	[0.5222, 0.5259]	4.1277	≪ .0001	0.0588
Indonesian	0.5264	0.0690	[0.5245, 0.5284]	0.5237	0.0675	[0.5218, 0.5256]	9.4220	≪ .0001	0.1343
Vietnamese	0.5214	0.0742	[0.5193, 0.5234]	0.5218	0.0705	[0.5199, 0.5238]	-1.2914	.1966	-0.0184
Turkish	0.5256	0.0675	[0.5237, 0.5275]	0.5230	0.0693	[0.5210, 0.5249]	7.7257	≪ .0001	0.1101
English	0.5281	0.0683	[0.5262, 0.5300]	0.5227	0.0701	[0.5207, 0.5246]	15.1285	≪ .0001	0.2156

### 3.3. Neural architecture

In the present experiment, in the light of the reduced dimensionality of the semantic with respect to the visual vectors, we constructed a model with an LSTM layer comprising 50 hidden units. The LSTM layer was followed by a dense layer with an equivalent shape. All the other hyperparameters were left unaltered with respect to the previous experiment.

### 3.4. Experimental conditions

We organized our experimental conditions following the same procedure as in Experiment 1. We divided the 24,612 concepts into a training (19,690 items) and a test set (4,922 items), with a 0.8 train-test split ratio. Then, we constructed six conditions where we trained the LSTM networks in the concatenation of the training set data in five languages and tested it in the test set relative to the language that was excluded from training. Once again, the models were tested in a data sample where the concepts, the semantic vectors, and the language were not represented in the training set. The results of the experimental models were compared with the ones achieved by a randomized baseline, trained on datasets where the correspondence between the phonetic vectors in input and the semantic vectors in output had been shuffled. All the models were trained on 98,450 samples and tested on 4,922 items.

### 3.5. Results

Table 3 reports the results of the models paired with their random counterparts. With the only exception of the transfer to Vietnamese, all the models outperformed their randomized baselines, with the contrast between the two models reaching statistical significance (although with marginal effect sizes). The above-chance performance of the cross-family networks is consistent with the hypothesis that a certain amount of cross-linguistic correspondence between form and meaning is stable across languages, and thus can be exploited when predicting a word's meaning in a previously unseen language. Moreover, this correspondence is not limited to visual vocabulary (Experiment 1) nor a subset of culture-independent concepts (Blasi et al., 2016; Wichmann et al., 2010) but can be captured at a lexicon-wide level. With respect to our previous experiment, the present results show that phonological patterns also

reflect a component of meaning that is encoded in language use, as reflected by a distributional semantic representation.

## 4. Experiment 3

In our previous experiments, we showed that our LSTM models were able to detect iconic cues in the input and exploit them when making predictions about the physical-geometrical properties of a word's referent and its language-based semantic representation. In our third experiment, we aimed to investigate whether the sound of a word encoded consistent cues about its syntactic behavior, expressed as its word class. It is widely accepted that word classes share relevant phonological properties *within* languages; nonetheless, the dominant view on systematicity in vocabulary holds that these properties fluctuate *across* languages (Dingemanse et al., 2015). We wanted to test the hypothesis that word classes might instead be organized according to consistent phonological principles, with some structural limitations on their cross-lingual variation. To address this issue, we relied on transfer learning in a classification setting, training an LSTM model to associate phonetic vector sequences with the hot-encoding of their word classes.

Since we did not modify the procedure for obtaining phonetic vectors from the letter strings in input, a description of the methodology can be found in Subsection 2.3. The experimental pipeline is depicted in Fig. 1 (dashed-dotted line).

### 4.1. Word-class representations

The word-class representations employed in the present experiment consisted of the one-hot encoding of the part-of-speech (PoS) tags of the vocabulary of the British National Corpus, released by Kilgarriff (1997). We collapsed the original 54 tags into 11 coarse supertags, which corresponded to the dimensions of the hot-encoded embedding. From the original database (208,656 items), we removed all the words with an ambiguous PoS tag (i.e., that were associated with more than one syntactic label); the resulting list comprised 152,855 items.

### 4.2. Translation

The words associated with a univocal PoS tag were translated into five languages with the combination of `word2word` and `fastText`, following the same procedure as in our previous experiments. The number of items for which a translation was available in all five languages was 24,246. No infinitive marker survived the translation procedure, so we removed the corresponding dimension from the vectors. With this procedure, we assumed an alignment of word classes across languages. This is a rather strong assumption, which is not likely to be fully supported by our data; however, different language-specific PoS taggers often classify words according to different tag sets, whereas our experimental procedure called for a shared classification schema.

Table 4  
Results of the transfer of the neural classifier

Language	Cross-lingual Model			Randomized Baseline			Contrast	
	Accuracy	Precision	F1	Accuracy	Precision	F1	$\chi^2$	$p$
Arabic	0.1001	0.4194	0.1396	0.0155	0.0009	0.0017	794.7239	$\ll$ .0001
Hungarian	0.1462	0.4018	0.1828	0.0168	0.0008	0.0014	1278.9005	$\ll$ .0001
Indonesian	0.1208	0.4419	0.1717	0.0183	0.2804	0.0049	971.0473	$\ll$ .0001
Vietnamese	0.0880	0.4060	0.1311	0.0159	0.0009	0.0017	622.6544	$\ll$ .0001
Turkish	0.1933	0.4481	0.2570	0.0286	0.3316	0.0056	1637.8704	$\ll$ .0001
English	0.1127	0.3977	0.1498	0.0360	0.3911	0.0233	503.9111	$\ll$ .0001

*Note.* We do not report recall scores since weighted recall is mathematically equivalent to accuracy in multi-class classification.

### 4.3. Neural architecture

Given that the experimental setting consists of a multiclass classification problem, categorical cross-entropy was used as the objective function, and for the same reason the softmax activation function was adopted for the output layer, and accuracy, precision, and F1 score were employed as metrics. After the removal of the infinitive markers, the vectors in the output were 10-dimensional arrays; thus, the output layer was equipped with 10 neurons. In the light of the reduced dimensionality of the word-class vectors in the output, we reduced the size of the LSTM layer to 25 units. All the remaining hyperparameters were left unaltered with respect to our previous experiments.

### 4.4. Experimental conditions

In order to construct our experimental conditions, we first split the 24,246 original items in a training set and a test set, which both included 12,123 instances. We employed a 0.5 split ratio in this experiment in order to have a reasonable number of instances in the test set for the minority classes: while upsampling is a useful procedure for balancing the classes in the training set, there is no use in upsampling the items in the test set. Thus, we dealt with class imbalance by randomly oversampling all the classes but the majority one in the training set, which reached 64,600 items. The usual six experimental conditions were devised by concatenating the training data in all the languages but one, and employing as test set the test data in the language that had been excluded from the training. The results of the experimental models were assessed by comparing their performances to the ones obtained by the parallel randomized baselines, where the order of the word-class vectors in output had been shuffled. All the models were trained on 323,000 instances and tested on 12,123 samples.

## 5. Results

Table 4 reports the results of the transfer of the LSTM-based classifiers. The first three columns indicate the accuracy, the weighted precision, and the weighted F1 score obtained



by the cross-lingual models, whereas the following three columns specify the same performance indexes relative to the randomized baseline models. The significance of the contrast between the accuracy of the parallel models was assessed by means of the McNemar test, a statistical test employed on paired nominal data. In all cases, the standard  $\chi^2$  calculation was employed, since the number of observations did not require us to resort to the exact binomial test. In all the experimental conditions, the cross-lingual models outperformed the randomized baselines by a wide margin in all the metrics considered, with all the contrasts reaching statistical significance.

### 5.1. Discussion

The results presented in the previous section are in line with our predictions and provide empirical evidence in favor of the existence of a universal phonetic substrate underlying word class distinctions across languages. To our knowledge, the present study constitutes the first attempt to refute the idea of a within-language idiosyncrasy in lexical systematicity through computational methodologies and at a large scale. We showed that the relationship between phonological profiles and word classes can be effectively transferred across language families, yielding language-independent generalizations in the mapping. Hence, systematicity should be regarded as a candidate universal feature underlying word formation. The view that iconic links are shared across languages should then be complemented by the finding that the phonological profiles of the lexical items are linked not only to their meaning but also to their organization in grammatical and distributional clusters.

### 5.2. Follow-up analyses

Once we verified that word classes are characterized by cross-linguistically stable phonological clusters, a natural question that arises is whether phonosyntactic information is uniformly distributed across syntactic categories, or whether some grammatical clusters incorporate stronger correspondences with their phonetic realization. To do so, we calculated the average accuracy of the cross-lingual models for each PoS in our tagset. The results are summarized in Fig. 3, which reports the accuracy aggregated by the PoS tag for each of the languages, as well as a second-order mean across all six languages considered in the study. Overall, our results are consistent across languages: the average pairwise correlation between the PoS-aggregated accuracy in all the combinations of two languages is  $r = .7738$ . The word class predicted with the highest accuracy by the cross-lingual models are interjections. This result is not surprising: Interjections directly express instinctive reactions (Bloomfield, 1984) and can be closely related to their spontaneous manifestation (Wharton, 2003); hence, it is natural to find a more transparent link between their phonoarticulatory expression and their class. Furthermore, this result is aligned with various findings documented in the literature. For instance, interjections are explicitly judged as the most iconic PoS by English speakers (Winter et al., 2017), and the interjection “Huh?” shows a particularly stable phonological realization, being found in roughly the same form in spoken languages across the world (Dingemanse et al., 2013). The accuracy in the other PoS does not seem to follow any clear pattern with respect to the linguistically relevant distinction between content and

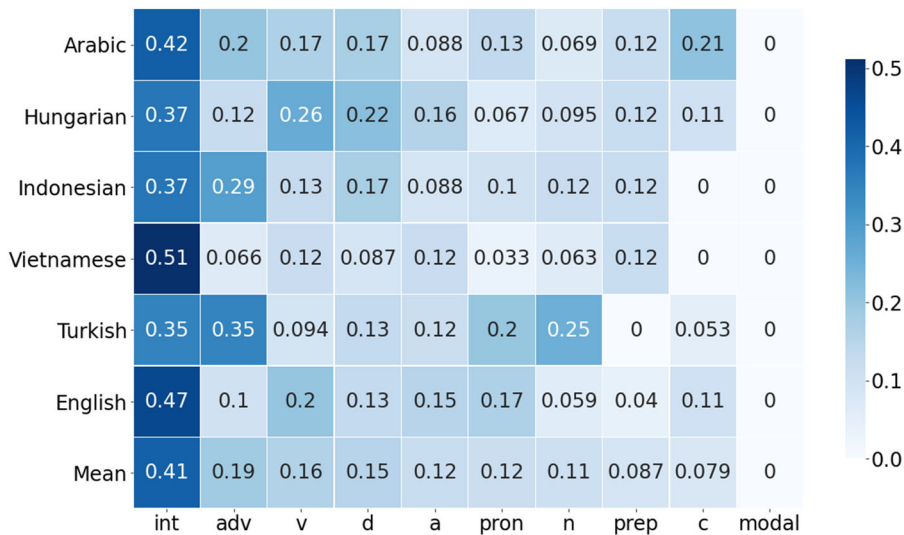


Fig. 3. Transfer accuracy averaged by PoS tags.

function words. For instance, adverbs are predicted with very high accuracy, but they can behave as both function and content words. Then, verbs, which are mostly content words with the exception of auxiliaries, occupy the third position in the scale, but are immediately followed by determiners. This suggests that the phonosyntactic clustering encoded in the phonological structure of the lexicon is subjected to a more subtle distinction than the coarse contrast between function and content words. The accuracy averaged by PoS that we obtained in the classification task is coherent with the cross-linguistic kinship of ideophones to other syntactic classes (Dingemanse, 2021). Ideophones are often connected to – or realized as – adverbs (for instance in Gbaya, as reported by Roulon-Doko, 2001), verbs (as in Shona, see Fortune, 1971), and adjectives (as in Ewe, see Ameka, 2001). These syntactic classes obtained comparably high-performance scores in our analysis, occupying the second, third, and fifth positions in our ranking. The high-performance scores obtained for determiners are reminiscent of the well-known role of iconicity in deictic demonstratives (Johansson & Zlatev, 2013; Johansson & Carling, 2015), where the pitch is associated with spatial distance (Ulta, 1978; Traunmüller, 1994; Woodworth, 1991). However, an important distinction must be made with respect to the correspondence between our findings and the ones we just reported. In Experiment 3, we showed that the words' sounds can be associated with their syntactic class, with this association varying in strength across word classes. The studies we summarized above showed instead that word sounds have a different association with their meaning as a function of their syntactic class. These two kinds of relationships are inherently different: The instances of a syntactic class can vary a lot in their meaning, and this is particularly clear for open lexical classes. Nonetheless, we showed that some word classes display a certain level of cross-linguistic phonetic regularity, and the same classes have been shown to exhibit a high internal consistency in the mapping between sound and meaning. Taken together, our results

and the findings presented in previous literature show an interesting convergence of systematic and iconic information, where syntactic classes with a more distinctive phonetic profile are also characterized by higher phonosemantic transparency. Our last study thus complements the body of findings demonstrating that iconic phonological patterns are not uniformly distributed in the lexicon, and shows that this asymmetry is mirrored by the way in which different languages encode grammatical classes through phonology.

## 6. Conclusion

This research effort constitutes a large-scale deep learning-based analysis of non-arbitrariness in language. Our key contribution consists in showing that the vocabulary is structured in partial resemblance to the visual world and contains phonological cues about the meaning of a word and its syntactic class.

Iconicity implies a meaningful link between form and meaning, such that the relationship between the two domains can be described as analogical. Unfortunately, our experimental approach does not ensure that the relationships learned by these models are analogical in nature. In fact, the general opaqueness of the deep learning methods we employed prevents us from understanding the kind of correspondences that have been exploited by the networks in order to make their predictions. For some of the anecdotal studies that we reported in the Introduction, the authors were able to interpret the link between the sound and meaning they detected. For instance, in the boubu-kiki experiment, the round vowel /u/ might suggest the presence of rounded shapes in its referent, in a quasi-synesthetic analogical relationship. Unfortunately, scaling up to nearly the whole lexica of six different languages makes the explanation of these correspondences an unrealistic enterprise, and the phonovisual, phonosemantic, and phonosyntactic biases that our networks exploited in their predictive behavior elude a precise explanation. However, it should be noted that the cross-lingual setup of our experiments mitigates the problem. If our experiments were performed within a single language, the finding that similar sounds express similar meanings could not be directly ascribed to iconicity, as it could be simply driven by the affixation of certain bound morphemes to the same roots. For instance, the word *clearly* is both semantically and phonologically close to the word *clear*, but this relationship is not dependent on sensorimotor analogies, but simply on the fact that the two words share the same root. On the other hand, if our experiments were performed on different languages belonging to the same language family, another possible source of non-arbitrariness that is not iconic in nature could be the etymological relatedness between words in typologically close languages. However, there are not many reasons that can account for a large-scale correspondence between sound and meaning that is detected across language families. First, this correspondence could be iconic in nature, as we have described it insofar. While the finding that iconic links are shared across different language families might be surprising at first, their cross-lingual consistency is nonetheless predicted by their own definition. Iconic biases are related to sound and meaning representations by means of perceptuomotor analogies (Dingemanse et al., 2015). Being grounded in the sensory and motor systems, which are relatively culture-invariant components of human

cognition, there are no reasons to expect them to display an ample degree of variation across languages. Closely related to the iconic account for the form-meaning association biases is the indexical explanation. Words can be related to their associated sensory experience by means of direct resemblance but also on account of their co-occurrence (Dingemanse, 2021). For instance, when looking at a perceptually pleasing landscape, one may utter the vocalization “wow.” This phonetic sequence does not bear any resemblance with the visual features of the landscape but is a typical *response* to an enchanting view, which routinely follows it. Cases of this kind might constitute a portion of the items in our second and third experiments (especially in the case of interjections), but we doubt they could have played a role in our first study. The image labels in the input were not typical responses to the presentation of those images, but their names; hence, this semiotic alternative is not likely to have been a major determinant of our results. Alternatively, the sound-to-meaning link could relate to some functional constraints in the interactional environments in which speakers of different languages communicate. Similar communicative environments might lead to the independent evolution of similar ways of referring to things, as similar physical environments lead to the development of similar body plans (see Dingemanse et al., 2013). A third possibility holds that similar words would display similar phonological patterns for being underpinned by a common genetic infrastructure. This view has been proposed within the literature on interjections (Müller, 1873; Sapir, 1921; although see Dingemanse et al., 2013) but is difficult to support at a lexicon-wide level: positing innateness for a wide variety of linguistic items would hardly be realistic given the timescale involved in language evolution (Dingemanse et al., 2013; Thompson, Smith, & Kirby, 2012). Hence, we believe that the only reasonable alternatives that can account for our findings are the interactional and iconic explanations. The present work is not sufficient to disentangle these hypotheses, and we leave to future research an empirical test of their predictions.

The difficulty in interpreting the nature of the correspondences learned by our models is not confined to our first two experiments on phonovisual and phonosemantic iconicity. The cross-lingual stability of the word-class cues that characterize terms with similar syntactic behavior challenges the assumption that systematicity should be regarded as an idiosyncratic linguistic feature; however, the reason for this phonological coherence within word classes is not unequivocal. Indeed, we can identify in this condition the same theoretical alternatives that we highlighted above. We speculate that this finding might point to the role of sensorimotor processing in shallow syntax. For the word-class cues to be relatively invariant across languages, they must be rooted in (or at least related to) other domains of the human cognitive system that show a certain degree of cross-cultural stability, and the perceptual and the motor systems are ideal candidates with that respect. Alternatively, a more functional approach would posit for word classes usually uttered in similar interactional contexts to display a certain degree of similarity in their phonological realization. For instance, we could expect word classes employed more often or learned earlier during language acquisition to display phonemes that are easier to produce and process. Again, a contrast between these two alternatives falls beyond the scope and the explanatory power of this study, and we leave their proper assessment to future research.

Taken together, our findings show that a remarkable amount of information is encoded in a word's sound: phonological profiles seem to contain cues concerning not only its meaning but also its syntactic behavior. A phonetic representation should not be seen as a formal and symbolic transposition of a word's meaning, but rather as an iconic pointer to its perceptual, semantic, and syntactic representation. While the present work highlighted the pervasiveness of linguistic iconicity, the decoupling of form and meaning must be recognized as a fundamental feature of language as well: without a certain degree of arbitrariness, it would not be possible to denote a potentially infinite set of concepts and their relationships (Lockwood & Dingemans, 2015). Arbitrary and iconic principles should be regarded as distinct properties of language (Sidhu & Pexman, 2018), with the complementary functions of detaching and grounding it to the sensorimotor experience.

## Acknowledgments

Open Access Funding provided by Università degli Studi di Milano-Bicocca within the CRUI-CARE Agreement.

## Conflict of interest

The authors declare no conflicts of interest.

## Notes

- 1 Following the Omniglot genealogical classification of languages at <https://omniglot.com/writing/langfam.htm>
- 2 Publicly available at <https://github.com/facebookresearch/MUSE>
- 3 Hot-encoding is a procedure employed in computer science to represent categorical data as vectors. The vector is a  $1 \times N$  matrix, where  $N$  corresponds to the number of classes; it consists of zeroes in all the cells, with the exception of a one in the cell corresponding to the class it has to mark.
- 4 Publicly available at <https://www.marekrei.com/projects/vectorsets/>
- 5 Publicly available at <http://www.kilgarriff.co.uk/bnc-readme.html#raw>
- 6 Adjective, adverb, conjunction, determiner, infinitive marker, interjection, modal verb, noun, pronoun, preposition, verb. The conversion table can be found in the code (see <https://github.com/Andrea-de-Varda/iconicity-deep-learning>).
- 7 We thank Reviewer #1 for bringing this matter to our attention.
- 8 We are grateful to Reviewer #2 for raising this issue.

## REFERENCES

- Abramova, E., & Fernández, R. (2016). 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, CA: Association for Computational Linguistics.
- Abramova, E., Fernández, R., & Sangati, F. (2013). Automatic labeling of phonesthemic senses. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, pp. 1696–1701). Austin, TX: Cognitive Science Society.
- Allott, R. (2001). *The natural origin of language: The structural inter-relation of language, visual perception, and action*. Knebworth, England: Able Publishers.
- Ameka, F. K. (2001). Ideophones and the nature of the adjective word class in ewe. *Typological Studies in Language*, 44, 25–48.
- 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* (pp. 76–93). Cambridge, England: Cambridge University Press.
- 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.
- Bloomfield, L. (1984). *Language*. Chicago, IL: University of Chicago Press.
- Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, 143(3), 263.
- 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.
- Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*, volume 9. Amsterdam, the Netherlands: John Benjamins Publishing.
- Cabrera, J. C. M. (2012). The role of sound symbolism in protolanguage: Some linguistic and archaeological speculations. *Theoria et Historia Scientiarum*, 9, 115–130.
- Chen, Y. C., Huang, P. C., Woods, A., & Spence, C. (2016). When “bouba” equals “kiki”: Cultural commonalities and cultural differences in sound-shape correspondences. *Scientific Reports*, 6, 26681.
- Choe, Y. J., Park, K., & Kim, D. (2020). word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3036–3045). Paris, France: European Language Resources Association.
- Chollet, F. et al. (2015). Keras. [https://keras.io/getting\\_started/faq/](https://keras.io/getting_started/faq/).
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. arXiv preprint arXiv:1710.04087.
- Croft, W. (2002). *Typology and universals*. Cambridge, England: Cambridge University Press.
- Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3), 615–627.
- Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., K &, Zeller, J. (2021). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841), 20200390.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR09)* (pp. 248–255). Piscataway, NJ: IEEE.
- Dingemans, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10), 654–672.

- Dingemanse, M. (2021). Ideophones. *Oxford Handbook of Word Classes*. Oxford, England: Oxford University Press.
- Dingemanse, M., Blasi, D., Lupyan, G., Christiansen, M., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19, 603–615.
- Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, 92(2), e117–e133.
- 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), e78273.
- Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. London, England: Oxford University Press.
- Fontana, F. (2013). Association of haptic trajectories to takete and maluma. In *International Workshop on Haptic and Audio Interaction Design* (pp. 60–68). Berlin, Germany: Springer.
- Fortune, G. (1971). Some notes on ideophones and ideophonic constructions in Shona. *African Studies*, 30(3-4), 237–258.
- Fryer, L., Freeman, J., & Pring, L. (2014). Touching words is not enough: How visual experience influences haptic-auditory associations in the “bouba-kiki” effect. *Cognition*, 132(2), 164–173.
- Gallace, A., Boschini, E., & Spence, C. (2011). On the taste of “bouba” and “kiki”: An exploration of word-food associations in neurologically normal participants. *Cognitive Neuroscience*, 2, 34–46.
- Graven, T., & Desebrock, C. (2018). Bouba or kiki with and without vision: Shape-audio regularities and mental images. *Acta Psychologica*, 188, 200–212.
- Günther, F., Marelli, M., Tureski, S. & Petilli, M. A. (2021). Vispa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. PsyArXiv. doi:10.31234/osf.io/n4dmq.
- Günther, F., Petilli, M. A., Vergallito, A., & Marelli, M. (2020). Images of the unseen: Extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research*. Advance online publication. <https://doi.org/10.1007/s00426-020-01429-7>
- 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.
- Haiman, J. (1985). Natural syntax. iconicity and erosion. *Cambridge Studies in Linguistics London*, (44), 1–285.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10), 1–24.
- Hirata, S., Ukita, J., & Kita, S. (2011). Implicit phonetic symbolism in voicing of consonants and visual lightness using Garner’s speeded classification task. *Perceptual and Motor Skills*, 113(3), 929–940.
- Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203(3), 88–97.
- 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*. Berlin, Germany: Walter de Gruyter.
- Johansohn, N., Anikin, A., & Aseyev, N. (2020). Color sound symbolism in natural languages. *Language and Cognition*, 12(1), 56–83.
- Johansson, N., & Carling, G. (2015). The de-iconization and rebuilding of iconicity in spatial deixis: An Indo-European case study. *Acta Linguistica Hafniensia*, 47(1), 4–32.
- Johansson, N., & Zlatev, J. (2013). Motivations for sound symbolism in spatial deixis: A typological study of 101 languages. *Public Journal of Semiotics*, 5(1), 3–20.
- Joo, I. (2020). Phonosemantic biases found in Leipzig-Jakarta lists of 66 languages. *Linguistic Typology*, 24, 1–12.
- Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135–155.
- Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
- Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

- Köhler, W. (1947). *Gestalt Psychology: An introduction to new concepts in modern psychology*. New York: Liv-eright.
- 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.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10), 1995.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1), 151–171.
- Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–38.
- Levinson, S. C., Stephen, C., & Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Lison, P., Tiedemann, J., & Kouylekov, M. (2018). OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan (pp. 1742–1748). Paris, France: European Language Resources Association (ELRA).
- Locke, J. (1847). *An essay concerning human understanding*. New York: Kay & Troutman.
- Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6, 1246.
- 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.
- Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752), 20170137.
- Magnus, M. (2013). A history of sound symbolism. *The Oxford handbook of the history of linguistics* (pp. 191–208). Oxford, England: Oxford University Press.
- Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 5188–5196). Piscataway, NJ: IEEE.
- Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman & Co.
- 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.
- McLean, B. (2021). Revising an implicational hierarchy for the meanings of ideophones, with special reference to Japonic. *Linguistic Typology*, 25(3), 507–549.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, 55(4), 259–305.
- 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.
- Müller, F. M. (1873). *Lectures on the science of language*, volume 2. London: Longmans, Green, and Company.
- Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Language is far less arbitrary than one thinks: Iconicity and indexicality in real-world learning and processing. *Journal of Cognition*, 4(1), 38.
- Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28(1), 225–252.



- Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114(2), 173–186.
- Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon*, 5(2), 205–230.
- Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, 214(3), 373–380.
- Pejovic, J., & Molnar, M. (2017). The development of spontaneous sound-shape matching in monolingual and bilingual infants during the first year. *Developmental Psychology*, 53(3), 581.
- Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1, 227.
- Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language*, 117, 104194.
- 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.
- Rei, M., & Briscoe, T. (2014). Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 68–77). London: Routledge.
- 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.
- Roulon-Doko, P. (2001). Le statut des idéophones en gbayà. *Typological Studies in Language*, 44, 287–302.
- Sapir, E. (1921). *Language: An introduction to the study of speech*. New York: Harcourt, Brace.
- Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3), 225.
- Sathian, K., & Ramachandran, V. S. (2019). *Multisensory perception: From laboratory to clinic*. Amsterdam, The Netherlands: Elsevier.
- Saussure, F. D. (1964). *Course of general linguistics (cours de linguistique générale, 1959)*. Second impression. ed. by Charles Bally and Albert Sechehaye. Wade Baskin (Trans.). London: Peter Owen.
- Shillcock, R., Kirby, S., & McDonald, S. (2001). Filled pauses and their status in the mental lexicon. In *Disfluency in Spontaneous Speech (DiSS'01), ISCA Tutorial and Research Workshop (ITRW), Edinburgh, Scotland, UK, August 29–31, 2001* (pp. 53–56). Baixas, France: ISCA Archive.
- Shinohara, K., & Kawahara, S. (2010). A cross-linguistic study of sound symbolism: The images of size. In *Annual Meeting of the Berkeley Linguistics Society*, volume 36 (pp. 396–410). Berkeley, CA: Berkeley Linguistics Society.
- Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, 25(5), 1619–1643.
- Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Slavova, V. (2019). Towards emotion recognition in texts—A sound-symbolic experiment. *International Journal of Cognitive Research in Science, Engineering and Education*, 7(2), 41–51.
- Smith, J. L. (2011). Category-specific effects. *The Blackwell companion to phonology*. (pp. 1–25). Wiley Online Library.
- Speed, L. J., Atkinson, H., Wnuk, E., & Majid, A. (2021). The sound of smell: Associating odor valence with disgust sounds. *Cognitive Science*, 45(5), e12980.
- Tamariz, M. (2008). Exploring systematicity between phonological and context-co-occurrence representations of the mental lexicon. *The Mental Lexicon*, 3, 259–278.
- Thompson, B., Smith, K., & Kirby, S. (2012). Cultural evolution renders linguistic nativism implausible. In *The Evolution of language* (pp. 557–558). Singapore: World Scientific.
- Traunmüller, H. (1994). Sound symbolism in deictic words. In *Tongues and texts unlimited: Studies in honour of Tore Jansson on the Occasion of His Sixtieth Anniversary* (pp. 213–234). Stockholm, Sweden: Stockholms Universitet.
- Ulat, R. (1978). Size-sound symbolism. *Universals of Human Language*, 2, 525–568.

- Vainio, L., Schulman, M., Tiippana, K., & Vainio, M. (2013). Effect of syllable articulation on precision and power grip performance. *PloS One*, 8(1), e53061.
- Werner, H. (1948). *Comparative psychology of mental development*. New York: International Universities Press.
- Werner, H. (2011). L'unité des sens. *Intellectica*, 55(1), 159–170.
- Wescott, R. W. (1971). Linguistic iconism. *Language*, 416–428.
- Wharton, T. (2003). Interjections, language, and the “showing/saying” continuum. *Pragmatics & Cognition*, 11, 39–91.
- Wichmann, S., Holman, E., & Brown, C. (2010). Sound symbolism in basic vocabulary. *Entropy*, 12, 844–858.
- Winter, B., Perlman, M., Perry, L., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies*, 18, 443–464.
- Woodworth, N. L. (1991). Sound symbolism in proximal and distal forms. *Linguistics*, 29, 273–299.
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.