Phonovisual Biases in Language: is the Lexicon Tied to the Visual World?

Andrea Gregor de Varda[1*], **Carlo Strapparava**[2]
[1]Center for Mind/Brain Sciences – University of Trento
[2]Fondazione Bruno Kessler – FBK
andreagregor.devarda@studenti.unitn.it, strappa@fbk.eu

Abstract

The present paper addresses the study of cross-linguistic and cross-modal iconicity within a deep learning framework. An LSTM-based Recurrent Neural Network is trained to associate the phonetic representation of a concrete word, encoded as a sequence of feature vectors, to the visual representation of its referent, expressed as an HCNN-transformed image. 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 randomized baseline; we show that such an imaginative network is capable of extracting language-independent generalizations in the mapping from linguistic sounds to visual features, providing empirical support for the hypothesis of a universal sound-symbolic substrate underlying all languages.

1 Introduction

How do words refer? How is a sound of the vocal tract mapped onto an object, event or state that is external to the mind? The conventional perspective on vocabulary structure, rooted in the structuralist tradition [Saussure, 1964], advocates an arbitrary origin of the lexicon, where words would have developed historically as a cultural and social product and passed along by tradition [Bloomfield, 1994]. However, the denial of any correspondence between linguistic sounds and their denotation is inherently problematic: seeing the lexicon as an arbitrary cultural product prevents any attempt of inquiry into its structure and its relationships with other aspects of human biology, irrevocably confining it outside of the scope of scientific explanation [Allott, 2001].

An alternative to arbitrariness is iconicity, the idea that phonemes can convey meaning *per se*, i.e. not only through contrastive relations with other sounds but also through their intrinsic sound qualities. While several modern linguistic theories reject iconic principles at the lexical level, they endorse them at the syntactic level with various degrees of explicitness, acknowledging a parallelism between the structure of language and various facets of the structure of experience [Levinson, 2000]. For instance, sequences of forms tend to correspond to sequences of occurrences [Bybee, 1985; Perniss *et al.*, 2010]: in the sentence “I studied, walked, and showered” the reader will typically assume that the three actions were performed in the order in which they were uttered. Widening our focus to non-arbitrariness in general, it is safe to state that most linguistic theoretical frameworks do not posit that languages can be structured without any accountable principle (see for instance Bickerton [2012], Chomsky [2002]). Nonetheless, the idea of a principled systematicity of any kind in the lexical domain is sometimes dismissed as a doubtful exception at best. As noted by Allott [2001], the tension between the syntactic schematicity and the alleged lexical randomness raises the question of how a functional interface between these domains is even possible.

The idea that phonemes can carry inherent meaning had a long-standing tradition in philosophy before being nearly eradicated by the Saussurean axiom of the arbitrariness of the sign (see Magnus [2013] for a historical introduction). This fascinating hypothesis was revived in the late 1920s by two remarkable studies, which evidenced a surprising link between the participants’ intuitions about a figure’s name and its size [Sapir, 1929] or its shape [Köhler, 1929]. In the first study, participants were asked to match two tables of different sizes with the non-words ‘mil’ and ‘mal’, showing that the phonetic sequence containing the phone [a] was four times more likely to be associated with the larger object than the one containing the phone [i]. In the second study, 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. The latter experiment on shape symbolism had a higher resonance than the former in the cognitive science community: several studies exploited slight variations in the shapes and the phonetic forms of the non-words, and repeatedly demonstrated the psychological reality of the so-called ‘maluma-takete’ effect [Köhler, 1947; Werner, 1948], or ‘bouba-kiki’ effect, referring to the stimuli used by Ramachandran and Hubbard [2001]. These research efforts paved the way to a number of inquiries that replicated the former phonovisual correspondences in various geocultural contexts [Bremner *et al.*, 2013; Ramachandran and Hubbard, 2001] and at different developmental stages [Maurer *et al.*, 2006; Ozturk *et al.*, 2013].

[*] Contact Author

The results of the studies that employed explicit matching tasks were complemented by a body of findings that succeeded in identifying an implicit cross-modal correspondence between phonological form and shape [Hung *et al.*, 2017; Sidhu and Pexman, 2017], suggesting that the aforementioned effects might be pre-semantic and presumably lexical, at least in part. Besides shape and magnitude symbolism, other visual attributes that showed a correspondence with the respective phonetic linguistic sign are colour [Johanssohn *et al.*, 2020] and lightness [Hirata *et al.*, 2011].

Iconicity is not limited to oral languages: a growing weight of evidence is disclosing analogies between the form of a considerable subset of signs and their referents [Emmorey, 2001; Kuhn, 2020; Schlenker and Lamberton, 2012]. Neither it is a privileged attribute of human languages, as it has been documented also in bee dances [Bermúdez, 2007] and primate calls [Burling *et al.*, 1993]. Recently, linguistic iconicity has gone from being a marginal – although appealing – matter to being integrated into broader theories of language evolution [Ramachandran and Hubbard, 2001], processing [Lockwood and Tuomainen, 2015] and acquisition [Asano *et al.*, 2015; Imai *et al.*, 2008]. Rejecting the assumption of a totally arbitrary mapping between *referens* and *referent* sensibly reduces the problem space of language emergence, establishing constraints on the consensus in word choice. Hence, the reviewed findings on phonosymbolic mappings could provide a pivotal clue for understanding the origins of proto-languages, suggesting that there may be natural constraints on the modalities in which objects are referred to through sounds. Furthermore, in the context of language learning, a propensity to induce an iconic referentiality of the linguistic sign alleviates both 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. Iconic links would then help children learn perceptually grounded semantic concepts, and discover structures across spoken and contextual input. Phonovisual correspondences have been shown to affect different cognitive faculties, such as memory [Ramachandran and Hubbard, 2001], categorization [Lupyan and Casasanto, 2015], and emotion recognition [Slavova and others, 2019]; moreover, they exert an influence on actional processes such as phonatory behaviour [Parise and Pavani, 2011], spatial navigation [Rabaglia *et al.*, 2016], and hand grip [Vainio *et al.*, 2013].

Iconic sound-to-referent relations have been disclosed in different perceptual modalities, such as haptic touch [Fryer *et al.*, 2014; Graven and Desebrock, 2018], kinesthesis [Fontana, 2013], and taste [Gallace *et al.*, 2011]. Nonetheless, vision seems to hold a privileged relationship with the phonetic representation of the *denotatum*: cross-modal correspondences in the olfactory-gustative modality do not seem to exhibit cross-cultural consistency [Bremner *et al.*, 2013], and phonotactile biases might be at least partially mediated by the role of visual imagery [Fryer *et al.*, 2014]. Hence, the present work aims to explore the multifarious topic of lexical iconicity with the physical attributes of the referent intended as its visual features.

Within the computational framework, the analysis of iconic biases has mainly followed two general trends [Gutiérrez *et al.*, 2016]: a localist approach, aimed at identifying some islands of non-arbitrariness in language [Abramova *et al.*, 2013; Sagi and Otis, 2008], and a global program, directed toward an assessment of the pervasiveness and systematicity of linguistic iconicity [Dautriche *et al.*, 2017; Monaghan *et al.*, 2014; Tamariz, 2008; Pimentel *et al.*, 2019]. Our work fits into the second trend, and aims to extend the previous findings through an exploration of iconic 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; those that have done so, have generally focused on a small set of concepts or words on a massively multilingual scale [Blasi *et al.*, 2016; Wichmann *et al.*, 2010] (although see Pimentel *et al.* [2019] and de Varda and Strapparava [2021] for lexicon-wide studies). Our work, in contrast, aims to perform an analysis on phonosensory (and specifically phonovisual) correspondences on a representative sample of concrete objects in a selected set of languages. We evaluated the performance of a Long Short-Term Memory network (LSTM) in associating phonetic vector sequences with visual vectors picturing their referents, reporting an above-chance performance of the model on an unseen language. To the best of our knowledge, this is the first cross-modal and cross-linguistic study exploiting deep learning methodologies to find an iconic relation between a linguistic sound and the physical-geometrical attributes of its denotation. This effort provides an example of rewarding contamination between methodological advances driven by artificial intelligence research, and theoretical questions stemming from philosophy, linguistics and psychology.

2 Methods

In the present study, an LSTM-based Recurrent Neural Network is trained to associate the phonetic representation of a word to the visual representation of its referent. Visual representations consist in the output of a pre-trained Hierarchical Convolutional Neural Network (HCNN) in response to a forward pass of a given image, whereas the phonetic features corresponding to each image's label are expressed as sequences of phonetic vectors in 22 dimensions. The experimental pipeline is summarized in the flowchart in Figure 1.

2.1 Dataset

Our analyses were performed on the THINGS database [Hebart *et al.*, 2019], a resource that comprises 26,107 high-quality naturalistic images depicting a set of 1,854 diverse object concepts. The concepts were sampled systematically from concrete picturable and nameable nouns in the American English language, and the corresponding images (12 or more for each concept) were extracted through a large-scale web image search and cropped to square size. Each item of the dataset was composed by an image and a corresponding label, that were preprocessed independently as described in the following subsections. In order to limit the

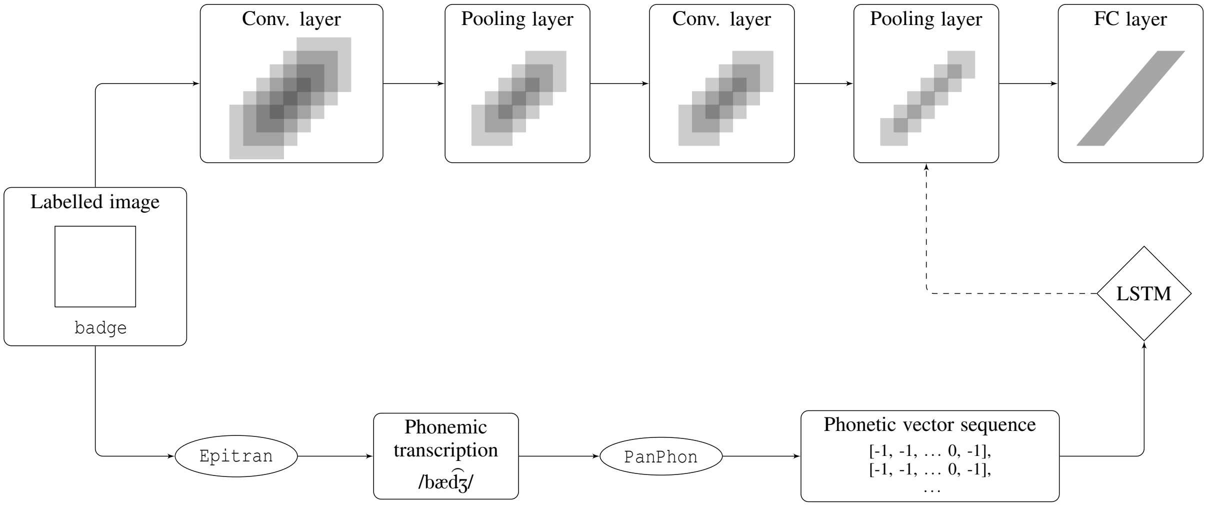

Figure 1: Schematic representation of the experimental pipeline. The upper stream reflects the processing stages of the image, whereas the inferior stream represents the phonetic vectorization of the letter string. For typographical reasons, only five layers of the VGG16 network are graphically depicted.

effect of morphological noise in the labels, we excluded from the database all the compound words before any subsequent analysis; with this exclusion criterion, the resulting dataset consisted of 22,268 images, depicting a set of 1,549 concrete words.

2.2 Translation

With the intention of maximizing the linguistic distance between the training and the test set, each image label in the dataset was translated in five languages belonging to five language families (see Table 1)[1]. In order to define a pipeline that could guarantee a high-quality translation for a sufficiently large amount of items, we first searched for lexical matches through word2word [Choe *et al.*, 2020], a collection of bilingual lexicons constructed from the publicly available OpenSubtitles2018 dataset [Lison *et al.*, 2018]; then, for the items for which a translation was not available through the aforementioned package, we employed the ground-truth bilingual dictionaries based on fastText, released by Facebook Research[2] [Conneau *et al.*, 2017]. We then included in the analyses only the words for which a translation was available in all the languages considered in the study – in other words, 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.

Language	Family	word2word	fastText	Missing
Arabic	Afroasiatic	84.38%	2.71%	12.91%
Hungarian	Uralic	82.18%	3.36%	14.46%
Indonesian	Austronesian	87.42%	1.48%	11.10%
Vietnamese	Austroasiatic	88.26%	0.06%	11.68%
Turkish	Turkic	81.15%	4.65%	14.20%
English	Indoeuropean	NA	NA	NA

Table 1: Languages, relative language families and translation data

2.3 Phonetic Representations

For each word in the multilingual 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 and 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 similarities would have been lost with a raw hot-encoding over the IPA vocabulary, along with the information-rich representation offered by a phonetic feature decomposition.

Before being loaded into the LSTM model, all the sequences were padded, with a maximum length of 15.

2.4 Visual Representations

The visual representations included in this study consisted of the outputs of the fifth max-pooling layer of the HCNN VGG16, a deep convolutional network for large-scale image recognition [Simonyan and Zisserman, 2015], in response to a forward pass of each image in the dataset. HCNNs con-

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

[2] Publicly available at https://github.com/facebookresearch/MUSE

sist of serial blocks of layers containing simple patterns repeated across the visual input. Despite the relative simplicity of every single layer, deep convolutional networks can compute complex transformations of the input data. We fed each image x in our stimulus set through the network to extract the resulting feature maps $\varphi(x)$ of `block5_pool`. The weights of the VGG16 network had been configured according to its pretraining on ImageNet. The outputs of the target layer were flattened before being processed by the LSTM model. We employed the output of the VGG16 network as a proxy of a representational format proper of the human visual processing system. Indeed, HCNNs have been employed to model different encoding steps of the mapping of visual stimuli to neural responses as measured in the brain, without being explicitly optimized to fit neural data [Yamins and DiCarlo, 2016].

2.5 Neural Architecture

An LSTM-based Recurrent Neural Network was trained to map the chains of phonetic feature vectors in input into the visual vectors in output. In contrast with standard feedforward neural networks, LSTMs are endowed with feedback connections, that allow them to process not only single data points, but also data sequences. The model was built with `Keras`, a deep learning framework for Python [Chollet, 2015]; it included a masking layer, followed by a single LSTM layer with 500 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 (25088) matching the dimensionality of the target visual vector, and equipped with the rectification nonlinearity (*ReLU*). Cosine similarity was used as both objective function and metric, and the Adam optimization method was employed for training [Kingma and Ba, 2014]. The hyperparameters were set without tuning.

2.6 Experimental Conditions

In defining the experimental conditions, we followed two main principles to limit to the furthest extent the effect of the etymological relatedness between the items in the training and the test set: we established for none of the images and none of the concepts to appear in both sets, and we applied the same constraint to the languages in which the concepts were translated. Therefore, we randomly divided the concrete concepts into two subsets with a train-test split ratio of 0.2. 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. We trained six different cross-lingual models with identical configurations following a non-random 6-fold cross-validation procedure. Each model was trained to associate the phonetic vector sequences corresponding to the label's translation in five languages to the respective visual vector in the training set; then, it was validated on the samples of the test set in the language that was excluded from the training. Thus, the train and test sets in all the experimental conditions were disjoint with respect to the concepts, the images representing them and the languages from which the phonetic vectorization was extracted. In order to define a baseline for the evaluation of each model's performance, we compared its results with the ones achieved by a parallel model, trained on a dataset where the correspondence between phonetic and visual vectors had been randomized; in practice, the mismatch between input and output vectors was obtained by randomly shuffling the order of the visual vectors in the training set. We will refer to this manipulation as the random condition. All models were trained on 66,985 samples (13,397 × 5 languages) and tested on 3,423 items.

3 Results

Table 2 reports the results of the cross-lingual models paired with their random counterparts. 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, as specified in the previous subsection. The following six columns of the table present the mean, the standard deviation and the 95% confidence intervals of the cosine similarity between the target visual vector and the cross-lingual (*cl*) 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 visual vector with the vectors generated by the two alternative models for each experimental condition. The eighth and ninth columns of the table present the t statistic and the associated p-value for each of the contrasts evaluated by the test. As a measure of effect size, we report in the last column of the table the Cohen's d relative to the standardized difference between the mean metric of the two alternative models.

Across all the experimental conditions, the cross-lingual models outperformed the randomized baselines, with all the comparisons reaching high statistical significance. Although there are considerable differences in the results of the alternative models reported in Table 2, the multilingual settings always yielded higher performance scores with respect to the randomized baselines. The confidence intervals of the *cl* models do not overlap with the intervals of the *r* models, suggesting that the phonovisual correspondences in the lexicon can be learned in any direction and generalized to all the languages included in our analyses. Additionally, the measures of effect size suggest a medium-to-large difference across models, with the only marginal exception of the zero-shot transfer to Vietnamese.

4 Discussion

The LSTM network trained on multilingual data showed the ability to induce cross-linguistic regularities in sound-to-vision mappings, suggesting that linguistic data alone contain the sufficient amount of information to encode for phonovisual biases. Hereby, we wish to clarify our intentions in using the two neural architectures for processing the images and their link to linguistic sounds. Our purposes were structurally different for the HCNN VGG16 and the LSTM: in the former case, we employed the convolutional neural network to transform the raw RGB images in input into cognitively inspired visual representations; hence, the neural network was

Language	**Cosine**$_{cl}$	**s**$_{cl}$	**95% CI**$_{cl}$	**Cosine**$_r$	**s**$_r$	**95% CI**$_r$	t	p	**d**
Arabic	0.2343	0.0411	[0.2329, 0.2357]	0.2243	0.0382	[0.2230, 0.2256]	10.42	3.01^{-25}	0.6600
Hungarian	0.2382	0.0410	[0.2368, 0.2396]	0.2278	0.0386	[0.2266, 0.2291]	10.78	6.83^{-27}	0.6321
Indonesian	0.2391	0.0394	[0.2378, 0.2404]	0.2243	0.0399	[0.2229, 0.2256]	15.45	6.13^{-53}	1.4385
Vietnamese	0.2320	0.0431	[0.2306, 0.2335]	0.2224	0.0384	[0.2211, 0.2237]	9.77	2.16^{-22}	0.4544
Turkish	0.2381	0.0404	[0.2367, 0.2394]	0.2257	0.0387	[0.2244, 0.2270]	12.99	3.94^{-38}	1.1956
English	0.2389	0.0418	[0.2375, 0.2403]	0.2228	0.0384	[0.2216, 0.2241]	16.60	1.01^{-60}	1.3220

Table 2: Results by experimental condition

intended as an approximation of the human perceptual system. In the latter case, we adopted the LSTM-based Recurrent Neural Network in order to uncover hidden correspondences between two different representational formats – a task that arguably requires the use of complex transformations.

While the behavioural studies presented in the Introduction disclosed a strong and consensual link between *meaningless* speech sounds and magnitude, colour and geometrical shape, we aimed to explore the link between *meaningful* speech sounds and visual representations. In the light of the evidence that highlights the role of visual imagery in shaping cross-modal phonosymbolic biases [Fryer *et al.*, 2014], we built an imaginative network trained to elicit a visual representation relying on a word's sound. The network was shown to learn a generational process where the images produced in response to the phonetic inputs resembled their actual referents more than what would be expected by chance. Our strict manipulation of the linguistic distance between the languages in the training and in the test set allows us to rule out the effect of any etymological relatedness between the different languages' vocabularies.

We believe that the central interest of our findings resides in the learning process being cross-linguistic, yielding generalizations that are language-independent. Indeed, crosslingualism is a crucial testbed for the purpose of identifying a universal sound-symbolic substrate underlying all languages, as opposed to language-specific idiosyncratic systematicity; whether this symbolic underpinning is to be considered innate [Ramachandran and Hubbard, 2001] and possibly reminiscent of a primordial state of neonatal synesthesia [Maurer and Mondloch, 2005] or implicitly induced by statistical correlations in shared sensorimotor experiences [Ernst, 2007; Spence, 2011] is beyond the scope and the explanatory power of the present study.

5 Conclusion and Future Directions

The present paper aims to contribute to the growing body of evidence against a purely cultural and arbitrary origin of the lexicon. Its purpose is to demonstrate that the vocabulary is profoundly entangled with the visual world, and that the correspondence between linguistic sounds and visual representations is a candidate universal feature underlying word-formation, shared across languages and language families. Moreover, it shows that this correspondence can be efficiently captured by a computational system in a selected subset of the lexicon. Crucially, the regularities detected by the network were not language- or item-specific, since in each experimental condition the items in the test set consisted of novel concepts, expressed in a previously unseen language and depicted by original images.

We wish to highlight the fact that the present work does not disclaim in any way the significance of a certain degree of arbitrariness as a fundamental property of language: without dissociating form and meaning it would not be possible to denote a potentially infinite set of concepts [Lockwood and Dingemanse, 2015]. Iconic and arbitrary principles are not mutually exclusive [Sidhu and Pexman, 2018]: on the contrary, they complement each other enriching the lexicon with their specific qualities.

Once the pervasiveness of lexical iconicity in the concrete vocabulary is acknowledged, a number of questions must be addressed to properly understand the phenomenon under scrutiny. Substantiating the claim of a cross-linguistic coupling between phonetics and vision was an imperative step, but it does not inform us on the *locus* of the biases that shape the cross-modal correspondences. We leave for future research an assessment of the precise visual attributes that show a privileged link with certain phonetic features, as well as an analysis of the relative contribution of shape, magnitude and colour to the phonovisual correspondences.

Acknowledgements

We thank Chiara Saponaro, Raffaella Bernardi and Stella Frank for their valuable insights, suggestions and comments.

References

[Abramova *et al.*, 2013] Ekaterina Abramova, Raquel Fernández, and Federico Sangati. Automatic labeling of phonesthemic senses. In *Proceedings of CogSci-2013*, 2013.

[Allott, 2001] Robin Allott. *The natural origin of language: the structural inter-relation of language, visual perception, and action*. Able Pub, 2001.

[Asano *et al.*, 2015] Michiko Asano, Mutsumi Imai, Sotaro Kita, Keiichi Kitajo, Hiroyuki Okada, and Guillaume Thierry. Sound symbolism scaffolds language development in preverbal infants. *Cortex*, 63:196–205, 2015.

[Bermúdez, 2007] José Luis Bermúdez. *Thinking without words*. Oxford University Press, 2007.

[Bickerton, 2012] Derek Bickerton. The origins of syntactic language. In *The Oxford handbook of language evolution*. 2012.

[Blasi *et al.*, 2016] Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823, 2016.

[Bloomfield, 1994] Leonard Bloomfield. *Language*. Motilal Banarsidass Publ., 1994.

[Bremner *et al.*, 2013] Andrew J. Bremner, Serge Caparos, Jules Davidoff, Jan de Fockert, Karina J. Linnell, and Charles Spence. "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, 126, 2013.

[Burling *et al.*, 1993] Robbins Burling, David F. Armstrong, Ben G. Blount, Catherine A. Callaghan, Mary Lecron Foster, Barbara J. King, Sue Taylor Parker, Osamu Sakura, William C. Stokoe, Ron Wallace, et al. Primate calls, human language, and nonverbal communication [and comments and reply]. *Current Anthropology*, 34(1), 1993.

[Bybee, 1985] Joan L. Bybee. *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing, 1985.

[Choe *et al.*, 2020] Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of LREC 2020*, 2020.

[Chollet, 2015] François *et al* Chollet. Keras. https://keras.io, 2015.

[Chomsky, 2002] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[Conneau *et al.*, 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[Dautriche *et al.*, 2017] Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8), 2017.

[de Varda and Strapparava, 2021] Andrea Gregor de Varda and Carlo Strapparava. A layered bridge from sound to meaning: Investigating cross-linguistic phonosemantic correspondences. In *Proceedings of CogSci-2021*, 2021.

[Emmorey, 2001] Karen Emmorey. *Language, cognition, and the brain: Insights from sign language research*. Psychology Press, 2001.

[Ernst, 2007] Marc O. Ernst. Learning to integrate arbitrary signals from vision and touch. *Journal of vision*, 7(5), 2007.

[Fontana, 2013] Federico Fontana. Association of haptic trajectories to takete and maluma. In *International Workshop on Haptic and Audio Interaction Design*, pages 60–68. Springer, 2013.

[Fryer *et al.*, 2014] Louise Fryer, Jonathan Freeman, and Linda Pring. Touching words is not enough: How visual experience influences haptic–auditory associations in the "bouba–kiki" effect. *Cognition*, 132(2):164 – 173, 2014.

[Gallace *et al.*, 2011] Alberto Gallace, Erica Boschin, and Charles Spence. On the taste of "bouba" and "kiki": An exploration of word–food associations in neurologically normal participants. *Cognitive neuroscience*, 2:34–46, 2011.

[Graven and Desebrock, 2018] Torø Graven and Clea Desebrock. Bouba or kiki with and without vision: Shape-audio regularities and mental images. *Acta Psychologica*, 188:200 – 212, 2018.

[Gutiérrez *et al.*, 2016] E. Dario Gutiérrez, Roger Levy, and Benjamin Bergen. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of ACL-2016*, Berlin, 2016.

[Hebart *et al.*, 2019] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 2019.

[Hirata *et al.*, 2011] Sachiko Hirata, Jun Ukita, and Shinichi Kita. Implicit phonetic symbolism in voicing of consonants and visual lightness using garner's speeded classification task. *Perceptual and Motor Skills*, 113(3):929–940, 2011.

[Hung *et al.*, 2017] Shao-Min Hung, Suzy J. Styles, and Po-Jang Hsieh. Can a word sound like a shape before you have seen it? sound-shape mapping prior to conscious awareness. *Psychological Science*, 28(3):263–275, 2017.

[Imai *et al.*, 2008] Mutsumi Imai, Sotaro Kita, Miho Nagumo, and Hiroyuki Okada. Sound symbolism facilitates early verb learning. *Cognition*, 109(1), 2008.

[Jakobson and Waugh, 2011] Roman Jakobson and Linda R. Waugh. *The sound shape of language*. Walter de Gruyter, 2011.

[Johanssohn *et al.*, 2020] Niklas Johanssohn, Andrey Anikin, and Nikolay Aseyev. Color sound symbolism in natural languages. *Language and Cognition*, 12(1), 2020.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[Köhler, 1929] Wolfgang Köhler. *Gestalt psychology. Liveright*. Oxford, England, 1929.

[Köhler, 1947] Wolfgang Köhler. *Gestalt Psychology: An introduction to new concepts in modern psychology*. Liveright, 1947.

[Kuhn, 2020] Jeremy Kuhn. Logical meaning in space: Iconic biases on quantification in sign languages. *Language*, 96(4):e320–e343, 2020.

[Levinson, 2000] Stephen C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[Lison *et al.*, 2018] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of LREC 2018*, Miyazaki, Japan, 2018.

[Lockwood and Dingemanse, 2015] Gwilym Lockwood and Mark Dingemanse. Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, 6:1246, 2015.

[Lockwood and Tuomainen, 2015] Gwilym Lockwood and Jyrki Tuomainen. Ideophones in japanese modulate the p2 and late positive complex responses. *Frontiers in psychology*, 6:933, 2015.

[Lupyan and Casasanto, 2015] Gary Lupyan and Daniel Casasanto. Meaningless words promote meaningful categorization. *Language and Cognition*, 7(2):167–193, 2015.

[Magnus, 2013] Margaret Magnus. A history of sound symbolism. *The Oxford handbook of the history of linguistics*, pages 191–208, 2013.

[Maurer and Mondloch, 2005] Daphne Maurer and Catherine J Mondloch. Neonatal synesthesia: A re-evaluation. In *Synesthesia: Perspectives From Cognitive Neuroscience*. Oxford University Press, 2005.

[Maurer *et al.*, 2006] Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322, 2006.

[Monaghan *et al.*, 2014] Padraic Monaghan, Richard Shillcock, Morten Christiansen, and Simon Kirby. How arbitrary is language? *Philosophical transactions of the Royal Society of London*, 369, 2014.

[Mortensen *et al.*, 2016] David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016*, Osaka, Japan, 2016.

[Ozturk *et al.*, 2013] Ozge Ozturk, Madelaine Krehm, and Athena Vouloumanos. Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of experimental child psychology*, 114(2):173–186, 2013.

[Parise and Pavani, 2011] Cesare V. Parise and Francesco Pavani. Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, 214(3), 2011.

[Perniss *et al.*, 2010] Pamela Perniss, Robin Thompson, and Gabriella Vigliocco. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227, 2010.

[Pimentel *et al.*, 2019] Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. Meaning to form: Measuring systematicity as information. In *Proceedings of ACL-2019*, Florence, Italy, 2019.

[Quine, 1960] Willard Quine. *Word and object: An inquiry into the linguistic mechanisms of objective reference*. John Wiley, Oxford, England, 1960.

[Rabaglia *et al.*, 2016] Cristina D. Rabaglia, Sam J. Maglio, Madelaine Krehm, Jin H. Seok, and Yaacov Trope. The sound of distance. *Cognition*, 152:141–149, 2016.

[Ramachandran and Hubbard, 2001] Vilayanur S. Ramachandran and Edward Hubbard. Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34, 2001.

[Sagi and Otis, 2008] Eyal Sagi and Katya Otis. Semantic glimmers: Phonaesthemes facilitate access to sentence meaning. In *The 9th conference on Conceptual Structure, Discourse Language*, Cleveland, OH, 2008.

[Sapir, 1929] Edward Sapir. A study in phonetic symbolism. *Journal of experimental psychology*, 12(3):225, 1929.

[Saussure, 1964] Ferdinand de Saussure. Course of general linguistics. *Trans. Wade Baskin. London.*, 1964.

[Schlenker and Lamberton, 2012] Philippe Schlenker and Jonathan Lamberton. Formal indices and iconicity in ASL. In Maria Aloni et al., editor, *Logic, Language and Meaning*, pages 1–11. Springer Berlin Heidelberg, 2012.

[Sidhu and Pexman, 2017] David M. Sidhu and Penny M. Pexman. A prime example of the maluma/takete effect? testing for sound symbolic priming. *Cognitive Science*, 41(7):1958–1987, 2017.

[Sidhu and Pexman, 2018] David M. Sidhu and Penny M. Pexman. Five mechanisms of sound symbolic association. *Psychonomic bulletin & review*, 25(5):1619–1643, 2018.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, San Diego, CA*, 2015.

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

[Spence, 2011] Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.

[Tamariz, 2008] M. Tamariz. Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3, 2008.

[Vainio *et al.*, 2013] Lari Vainio, Mirjam Schulman, Kaisa Tiippana, and Martti Vainio. Effect of syllable articulation on precision and power grip performance. *PloS one*, 8(1):e53061, 2013.

[Werner, 1948] Heinz Werner. *Comparative psychology of mental development.* Follett Pub. Co., 1948.

[Wichmann *et al.*, 2010] Søren Wichmann, Eric Holman, and Cecil Brown. Sound symbolism in basic vocabulary. *Entropy*, 12, 2010.

[Yamins and DiCarlo, 2016] Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.