Phonological Layers of Meaning: A Computational Exploration of Sound Iconicity

Andrea Gregor de Varda
Centre for Mind/Brain Sciences
University of Trento
andreagregor.devarda@studenti.unitn.it

Carlo Strapparava
Fondazione Bruno Kessler (FBK)
strappa@fbk.eu

Abstract

The present paper aims to investigate the nature and the extent of cross-linguistic phonosemantic correspondences within a computational framework. An LSTM-based Recurrent Neural Network is trained to associate the phonetic representation of a word, encoded as a sequence of feature vectors, to its corresponding semantic representation in a multilingual vector space. The processing network is tested, without further training, in a language that does not appear in the training set. The performance of the multilingual model is compared with a monolingual upper bound and a randomized baseline. After the quantitative evaluation of its performance, a qualitative analysis is carried out on the network's most effective predictions, showing an inhomogeneous distribution of phonosemantic information in the lexicon, influenced by semantic, syntactic, and pragmatic factors.

1 Introduction

The idea of a consistent relationship between sound and meaning has held a particular fascination over philosophers and linguists (Plato, 1998). However, in recent times, this charming hypothesis has progressively lost the interest of scholars, especially in the post-Saussurean linguistic tradition, which emphasized the arbitrariness in such relation. The idea that sounds have inherent meanings has recaptured its original attractiveness in the field of cognitive sciences, where the attention has initially focused on the link between sound and shape. A prominent example of these naturally biased mappings came from Köhler's (1929) finding that, when asked to match two novel shapes with the non-words 'maluma' and 'takete', English-speaking adults tended to label as 'maluma' the curled shape, and as 'takete' the sharp one. This germinal study paved the way to several replications and expansions of its findings, that reproduced Köhler's results in different geocultural contexts (Bremner et al., 2013) and at different developmental stages (Maurer et al., 2006). Since then, different studies have tackled the topic of iconicity in language from a broader perspective, showing that adults can associate visually presented characters (Koriat and Levy, 1977) and auditorily presented words (Berlin, 1995) of a foreign language to their meaning, with an accuracy above chance.

Recently, linguistic iconicity has gone from being a marginal – although appealing – matter to being integrated into broader theories of language evolution and acquisition. Indeed, rejecting the assumption of an arbitrary mapping between sound and meaning sensibly reduces the problem space of language emergence, establishing constraints on the consensus of word choice. Furthermore, an iconic relation between a sound and its referent might help with memory consolidation in the process of language acquisition (Sathian and Ramachandran, 2019). Ramachandran and Hubbard (2001) speculate that phenomena as the one reported by Köhler might arise from neural connections among adjacent cortical areas, where the visual features of the referent, the appearance of the speaker's lips and the kinaesthetic features of the articulation are combined. According to their view, such neural connections would have influenced both the phylogenetic evolution and the ontogenetic development of language. Although the previous findings are consistent with this hypothesis, an alternative explanation must be taken into account: the roots of these correspon-

dences could be grounded in the knowledge of language, that allows children and adults to generalize the regularities in sound-to-meaning mappings from their native language to nonsense and foreign words. Under this rationale, phonosemantic relations would be implicitly learned from general recurrences in already known languages. A crucial aspect of this account lies in the fact that it does not posit any preexisting disposition wired in the human brain, moving the *locus* of linguistic iconicity from the mind to language itself. A natural question that arises from this perspective is whether linguistic information alone is sufficient to give rise to the phonaesthetic biases presented in the literature. A computational exploration of the phenomenon under scrutiny is a feasible way to approach the subject. The idea that phones have inherent meanings is relatively understudied within the computational framework, and most of the studies addressing the topic have either focused on a single language (Gutiérrez et al., 2016; Sagi and Otis, 2008; Abramova et al., 2013; Monaghan et al., 2014; Tamariz, 2008) or on a small set of concepts on a massively multilingual scale (Blasi et al., 2016; Wichmann et al., 2010). Surprisingly, no study to our knowledge has tackled the topic through a deep learning methodology, and no cross-linguistic investigation has been performed on a lexicon-wide level. The purpose of the present study is two-fold: first, we wish to explore the idea of a cross-linguistic correspondence between the phonetic and the semantic representation of a word on the whole lexicon, without any theory-driven restriction guiding our choice of the lexical items. Then, we aim to examine whether the meaning that is rooted in the sound that words are made of is homogeneously distributed in the lexicon. Ultimately, these two goals converge toward the research question hinted above, namely, whether linguistic information alone could suffice for the extrapolation of the phonosemantic biases reported in the present section. A possible way to answer this question is to assess the ability of a *tabula rasa* neural network to extend the regularities captured in a set of given languages to a previously unseen one. Although equipped with clear structural priors, neural networks do not conceal biases that resemble those assumed to model the aforementioned phonosemantic correspondences. If a processing network showed the ability to induce cross-linguistic regularities in sound-to-meaning mappings, this would suggest that linguistic data contain a sufficient amount of information to encode for phonosymbolic biases.

The present study aims to explore the possibility of a certain degree of cross-linguistic correspondence between sound and meaning that is already encoded in language. A Long Short-Term Memory network (LSTM) is trained on four languages to associate the sequence of sounds that compose a word, encoded as phonetic vectors, to its corresponding semantic representation in a multilingual vector space. Then the processing network is tested, without further training, on a language that does not appear in the set of languages on which the training has been performed. The performance of the multilingual model is compared with the results of (a) a monolingual model, trained and tested on different subsets of a single language's vocabulary, and (b) a baseline model, where the output vectors in the training are randomly shuffled. After the quantitative evaluation of its performance, a qualitative analysis is carried out on the network's most effective predictions.

2 Methods

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

2.1 Semantic vectors

The semantic representations included in the model, provided by Facebook Research, consist in multilingual word embeddings generated with `fastText` from Wikipedia data (Bojanowski et al., 2017) and aligned in a common vector space through a fully unsupervised methodology (Conneau et al., 2017)[1]. The present study is conducted on Italian, German, French, Vietnamese, and Turkish embeddings.

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

2.2 Phonetic vectors

For each word in the embedding dataset, we obtained its phonemic transcription with `Epitran`, a Python library for transliterating orthographic text in the International Phonetic Alphabet (IPA) format. Then, we converted the IPA string into a sequence of feature vectors in 22 dimensions with `PanPhon`, a package that traduces IPA segments into subsegmental articulatory features (Mortensen et al., 2016). It has been shown that phonologically aware models built on the linguistically motivated and information-rich representations yielded by the Epitran-PanPhon pipeline outperform the raw hot-encoding of character-based models in different tasks (Mortensen et al., 2016; Bharadwaj et al., 2016).

2.3 Neural architecture

An LSTM-based Recurrent Neural Network is trained to map the sequences of phonetic feature vectors in input into semantic vectors in output. The model is built with `Keras`, a deep learning framework for Python (Chollet et al., 2015); it includes a single LSTM layer with 172 units, a dropout of 0.2 and a recurrent dropout of 0.2. Cosine similarity is used as both objective function and metric, and the Adam optimization method is employed for training (Kingma and Ba, 2014). We adopted the *tanh* activation function for the output layer since its codomain corresponds to the range (-1, 1), in which the semantic vectors are defined. The hyperparameters are set without tuning.

2.4 Experimental conditions

The experimental conditions are characterized by different combinations of training and testing sets. In the multilingual condition, the model is trained for one epoch on the Italian, German, French, and Vietnamese datasets, and then tested in Turkish. Our unique concern in the language selection was that none of the languages in the training set was typologically close with the language presented in the test set. Turkish has been chosen for the test set since it is not considered to be related to any of the languages presented in the training set, at least within a reasonable time window. Indeed, Turkish is a Turkic language, whereas Italian, German and French are Indoeuropean, and Vietnamese belongs to the Austroasiatic language family. To establish a baseline for the evaluation of the model's performance, we trained a model randomly shuffling the output vectors. We will refer to this manipulation as the random condition. In the monolingual condition, which defines the upper bound of the network's performance, the LSTM is trained and tested on different subsets of the Italian dataset, with a train-test split ratio of 0.2. In order to compensate for the different dimensions of the training set (roughly one fifth of the multilingual sample), the monolingual model is trained for five epochs.

3 Results

Table 1 lists the test results for each of the models described in Section 2.4. The number of lexical items included in the training and in the test set are reported in the $\text{Dim}_{\text{train}}$ and the Dim_{test} columns, respectively. The last column of the table presents the average cosine similarity between the target semantic vector and the model's prediction for every word in the test set. As reported below, the multilingual model outperforms the random baseline, with a 0.0351 points higher average cosine similarity. As expected, the monolingual performance is stronger than the one achieved by the multilingual model, with a difference of 0.0453 in the metric. The relatively modest magnitude of the difference between the monolingual and the multilingual results should be attributed to the limited size of the training set in the former condition: increasing the number of epochs might have partially compensated for the shortage in the training data, but additional forward and back propagation on the same data might not be as effective as further training on unseen data, especially in terms of generalization. The general pattern of results, with the multilingual performance almost halfway between the monolingual and the random results, is in line with our predictions. The difference between the multilingual and the random condition is consistent with the hypothesis that a certain degree of cross-linguistic correspondence between phonetic and semantic representations is already encoded in language; moreover, it shows that, with sufficient training, this correspondence can be efficiently captured by an LSTM network.

4 Qualitative analysis

As previously mentioned, the LSTM network trained on multilingual data showed the ability to induce cross-linguistic regularities in sound-to-meaning mappings, suggesting that linguistic data

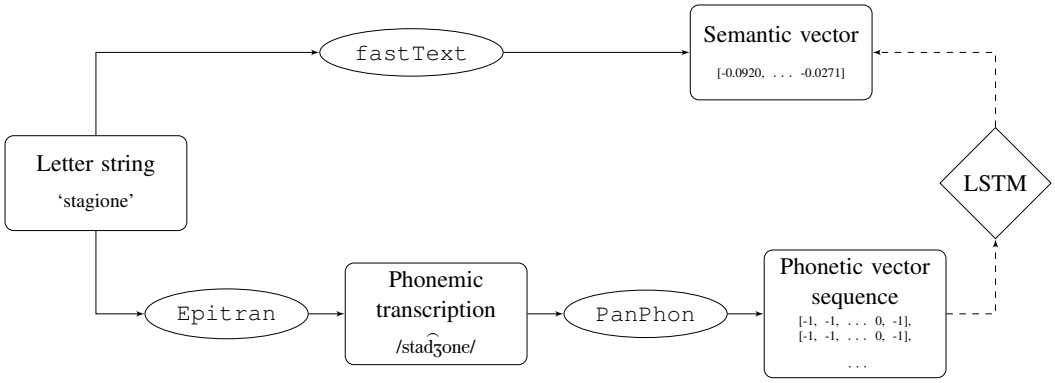

Figure 1: Schematic representation of the experimental pipeline

Model	**Dim$_{train}$**	**Dim$_{test}$**	**Cosine similarity**
Multilingual	794,870	199,108	0.4467
Random	794,870	199,108	0.4116
Monolingual	159,621	39,906	0.4920

Table 1: Test results by experimental condition

alone contain the sufficient amount of information to encode for phonosymbolic biases.

In the light of the results presented above, a natural question that arises is whether phonosemantic information is uniformly distributed in the lexicon, or some semantic areas tend to incorporate stronger correspondences with their phonetic counterparts. We hypothesized that some areas of the semantic space might show a more consistent mapping with their phonological realization, but without any clear *a priori* expectation on the regions that could reveal higher phonosemantic transparency, we addressed this problem through a data-driven qualitative analysis. We extracted from the test results of the multilingual condition 39,821 items (20% of the total), selecting the words that the network had predicted with the higher precision – that is, the words whose vector prediction had the higher cosine similarity with respect to the target. Then, we restricted the analysis by excluding the items with low frequency. We conjectured that it would be unlikely for rare and unfamiliar terms to convey phonosemantic relations without being etymologically related to other languages. For instance, across different disciplines, the technical jargon – whose instances are typically infrequent in corpora – is commonly derived from Greek and Latin roots. We employed the Twitter-based Turkish frequency estimates from the Worldlex dataset[2], that has been shown to outperform traditional frequency estimates in predicting lexical decision reaction times, thus exhibiting a higher cognitive validity (Gimenes and New, 2016). From the previously extracted items, we excluded those that were not in the list of the 20,000 most frequent words (that is, the 1.21% of the words with higher frequency). The resulting items were translated into English with `Googletrans`, a Python library that implements Google Translate API. The results of the analysis are reported in Table 2, where the items that satisfy the aforementioned constraints (from now on, the *quality subset*) are grouped into four intuitive categories according to their meaning and their grammatical function.

The most represented categories of words in this subset of efficiently predicted items are proper names and lexical borrowings, with the former generally associated with a higher cosine similarity between target and prediction. They are not reported in Table 2, since their detailed analysis is not relevant for the purposes of the study. However, the predominance of proper names over lexical borrowings is compatible with one of the basic postulates of model-theoretic semantics. It is generally assumed that proper names, unlike definite descriptions and generalized quantifiers, directly refer to entities in the world (Delfitto and Zamparelli, 2009); hence, they are expected to hold their exact meaning across languages.

The cross-linguistic consistencies in proper names and lexical borrowings are clearly due to contact between languages. Other word categories strongly associated with the phonosemantic fea-

[2]Publicly available at `http://worldlex.lexique.org`

Internal states	istediğimde (‘I want’), düşüncelerimi (‘my thoughts’), isteyenlere (‘those who want’), düşünsenize (‘imagine’), düşüncem (‘I thought’), açıkçası (‘frankly’), aşıksın (‘you are in love’), kendimde (‘in myself’)
Function words	vee (‘and’), kendileri (‘themselves’), onların (‘they’), gerektiğinde (‘when’), mıydın (‘did you’)
Interjections	hee (‘ooh’), boku (‘shit’), himm (‘uhm’)
Other	yaklaşım (‘approach’), poğaça (‘pastry’), demis (‘said’), gerçekmiş (‘real’), tabiiki (‘of course’), uygulamaları (‘applications’), gani (‘abundant’)

Table 2: Intuitive clustering of the model’s best predictions

tures detected by the network are undoubtedly more relevant in revealing lexical clusters with privileged sound-to-meaning mappings. For instance, a conspicuous portion of items in the quality subset is semantically linked to different internal states, with a predominance of concepts related to mental processes. The quality subset comprises also various function words (conjunctions, pronouns, and one auxiliary verb). This result is particularly informative since function words, being a closed-class category, are not as numerous as content words; therefore, their number of instances in the training set was most likely limited. An additional cluster in the quality set comprises three interjections, including one imprecation. Interjections express spontaneous feelings or reactions (Bloomfield, 1984) and can be closely related to their natural manifestation (Wharton, 2003); hence, it is not surprising to find a more transparent link between their phonoarticulatory expression and their meaning. Moreover, this result is consistent with the findings of Dingemanse et al. (2013), that show that the interjection “Huh?” is a universal, found in roughly the same form and function in spoken languages across the globe.

The present findings suggest that phonosemantic information is not uniformly distributed in the lexicon: the consistency of the mapping between sound and meaning seems to be influenced by semantic, syntactic, and pragmatic factors. Indeed, the semantic neighbourhood linked to internal states shows a privileged relationship between sound and meaning, whereas on the syntactic side function words seem to be favoured, if their absolute prevalence in the lexicon is taken into account. Moreover, interjections, which are characterized by a strong pragmatical valence, stand among the items predicted with the highest precision by the model.

5 Limitations and further directions

From a methodological standpoint, the reliability of the present results could benefit from the exclusion of lexical borrowings and proper names from the training and the test sets. Excluding etymologically related terms could further improve the reliability of the results, but at the costs of raising the difficulty of assessing the words’ relatedness in different languages, with the subsequent need of a proper metric.

Another confound that we wish to address in future research is the role played by morphological factors in aiding the cross-linguistic feature extraction performed by the network. FastText vectors exploit information related to subword character strings, and might therefore encode regularities pertaining to recurrent morphemes in the non-isolating languages in our dataset (Italian, German, French, and Turkish). We acknowledge that the network might have captured the recurrences encoded in the semantic vectors comprising the training set and their relationships with the corresponding phonetic feature vectors; indeed, we believe that this regularities might have played a relevant role in the monolingual condition, where the model might have learnt that morphologically related words (i.e. in this context, words that are similar at the character- and phoneme-level) tend to be associated with close subregions of the semantic space. Nonetheless, we do not see how this information could have altered significantly the performance in the multilingual condition. That said, we leave for future research an assessment of the algorithm’s performance on semantic vectors which lack access to subword-related information, such as `word2vec` (Mikolov et al., 2013), and in languages with opaque orthography (e.g. English and French) and non-concatenative morphology (e.g. Chinese)[3].

[3]We gratefully thank an anonymous reviewer for drawing our attention to this matter and suggesting the mentioned options to address this confound.

As for all the studies that employ artificial neural networks to draw conclusions on human cognition, it is mandatory to clarify some limitations on the extent of the inferences that can legitimately follow the presented results. The finding that a neural network can succeed in a task without the structural priors postulated in the human mind does not necessarily imply that these priors are not actually encoded in the brain: the assumption of a functional equivalence between artificial and biological processes needs to be independently motivated. Moreover, it should be noticed that the participants of the behavioural studies presented in Section 1 were not necessarily polyglots, whereas the promising cross-linguistic performances described in the results have been obtained with a multilingual model. In addition to these intrinsic methodological limitations, an account that does not assume any prior specification for the linguistically encoded phonosymbolic mappings would leave an open question concerning their origin. Hence, the present study does not claim to reject the multi-sensory integration hypothesis presented in the Introduction. Its purpose is simply to show that, in principle, linguistic information alone could suffice for a generalization in sound-to-meaning mappings.

References

Ekaterina Abramova, Raquel Fernández, and Federico Sangati. 2013. Automatic labeling of phonesthemic senses.

Brent Berlin. 1995. *Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature*, page 76–93. Cambridge University Press.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

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

Leonard Bloomfield. 1984. *Language*. University of Chicago Press. Google-Books-ID: 87BCDVsmFE4C.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Andrew J. Bremner, Serge Caparos, Jules Davidoff, Jan de Fockert, Karina J. Linnell, and Charles Spence. 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.

François Chollet et al. 2015. Keras. `https://keras.io`.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data.

Denis Delfitto and Roberto Zamparelli. 2009. *Le strutture del significato*. Itinerari Linguistica. Mulino, Bologna. OCLC: 695640183.

Mark Dingemanse, Francisco Torreira, and N. J. Enfield. 2013. Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE*, 8(11).

Manuel Gimenes and Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48(3):963–972.

E. Dario Gutiérrez, Roger Levy, and Benjamin Bergen. 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)*, pages 2379–2388, Berlin, Germany. Association for Computational Linguistics.

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

Wolfgang Köhler. 1929. *Gestalt Psychology*. Liveright.

Asher Koriat and Ilia Levy. 1977. The symbolic implications of vowels and of their orthographic representations in two natural languages. *Journal of Psycholinguistic Research*, 6(2):93–103.

Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. 2006. The shape of boubas:

sound–shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–322.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Padraic Monaghan, Richard Shillcock, Morten Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 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*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Plato. 1998. *Cratylus*. Hackett Publishing Company.

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

Eyal Sagi and Katya Otis. 2008. Semantic glimmers: Phonaesthemes facilitate access to sentence meaning.

K. Sathian and V. S. Ramachandran. 2019. *Multisensory perception: from laboratory to clinic*. Elsevier.

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

Tim Wharton. 2003. Interjections, language, and the 'showing/saying' continuum. *Pragmatics & Cognition*, 11:39–91.

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