Discovering Phonesthemes with Sparse Regularization

Nelson F. Liu[1,2], **Gina-Anne Levow**[2], **and Noah A. Smith**[1]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2]Department of Linguistics, University of Washington, Seattle, WA, USA
{nfliu,nasmith}@cs.washington.edu, levow@uw.edu

Abstract

We introduce a simple method for extracting non-arbitrary form-meaning representations from a collection of semantic vectors. We treat the problem as one of feature selection for a model trained to predict word vectors from subword features. We apply this model to the problem of automatically discovering phonesthemes, which are submorphemic sound clusters that appear in words with similar meaning. Many of our model-predicted phonesthemes overlap with those proposed in the linguistics literature, and we validate our approach with human judgments.

1 Introduction

Linguists have long held that language is arbitrary, or that a word's phonetic and orthographic forms have no relation to its meaning (de Saussure, 1916). For example, there is nothing about an apple that suggests that *apple* is the proper word for it—this link between meaning and the representation in language is arbitrary. Arbitrariness is a defining feature of human language, and it is a key component of the design features of language proposed by Hockett (1960).

Despite this, work over the last decades has revealed several exceptions to the arbitrariness of language. One such exception is iconicity, where the form of a word directly resembles its meaning. For example, Ohala (1984) showed that speakers tend to associate vowels with high acoustic frequency with smaller objects, while vowels with low acoustic frequency are associated with larger objects. In this case, speakers make a link between the phonetic form of a word and its perceived meaning because of an innate belief that smaller entities emit higher-frequency vowels while larger entities tend to emit low-frequency vowels.

Similarly, Köhler (1929) and Ramachandran and Hubbard (2001) observed a non-arbitrary connection between the shapes of objects and speech sounds. American college undergraduates and Tamil speakers were presented with a jagged shape and a rounded shape and asked which is *"kiki"* and which is *"bouba"*. In both groups, 95% to 98% selected the jagged shape as *"kiki"* and the rounded shape as *"bouba"*, demonstrating that the human brain connects sounds to shapes in a consistent way. D'Onofrio (2014) posits that the rounded shape is commonly named *"bouba"* since the mouth forms a rounded shape in producing the word, whereas pronouncing *"kiki"* requires a tighter, more angular mouth shape that seems more apt for the jagged object. In this case, there is a strong, non-arbitrary link between the articulatory properties of the sound and their perceived meaning.

Phonesthemes are another exception to the arbitrariness of language. Phonesthemes are non-compositional, submorphemic phonetic units that consistently occur in words with similar meanings. For example, the word-initial *gl-*, occurs at the beginning of many English words relating to light or vision, like *glint*, *glitter*, *gleam*, *glamour*, etc. (Hutchins, 1998; Bergen, 2004). The work of Hutchins (1998) includes a compilation of 46 phonesthemes proposed by linguists.

There is a body of previous work suggesting that phonesthemes are units in the mental lexicon of native speakers. For example, the work of Hutchins (1998), Magnus (2000), and Bergen (2004) uses priming experiments and other methods from psycholinguistics to demonstrate that phonesthemes significantly affect native speaker reaction times in a range of language processing tasks. In another line of work, Otis and Sagi (2008) and Abramova and Fernández (2016) verify phonesthemes by analyzing whether the words containing a given phonestheme are more semantically similar than expected by chance, where se-

mantic similarity is derived from a distributional semantic model.

While there has been much work in verifying previously proposed phonesthemes, there has been little work on automatically discovering new ones. In this work, our goal is to identify the likely phonesthemes of a language from a collection of semantic vectors. We do this by identifying the character or phoneme sequences that are predictive of word meaning by training a model to predict word vectors from subword features. Then, we use standard feature selection techniques to find a subset of features that best predict the vectors; this subset of features contains the model-predicted phonesthemes. Lastly, we validate the model-predicted English phonesthemes with human judgments and also find that many of our predicted phonesthemes overlap with those documented in previous work.

2 Method

To extract phonesthemes from a set of vectors, we want to find submorphemic units (e.g., character or phoneme n-grams) that are highly predictive of word meaning. We approach this problem through the lens of feature subset selection: given a model capable of predicting semantic vectors from submorpheme information, our goal is to select the subset of submorphemes (model features) that are most predictive. Intuitively, if a submorpheme is especially predictive of the word vectors, then it may be a meaning-bearing phonestheme.

We use linear regression to predict word vectors from binary feature vectors that encode the submorphemes occurring in a surface form. We use sparse regularization to select relevant features from this model, which enables it to automatically choose a subset of the submorpheme features that predict the vectors (our predicted phonesthemes).

Specifically, we regularize our linear regression model with the elastic net (Zou and Hastie, 2005). We used `scikit-learn` (Pedregosa et al., 2011) to train our models, and we tune the L_1 and L_2 regularization strengths on held-out error in 5-fold cross-validation.

Mitigating the Effect of Morphemes A principal concern is that the model will detect morphemes rather phonesthemes. Many past studies on the relationship between form and meaning in language (Shillcock et al., 2001; Monaghan et al., 2014; Gutiérrez et al., 2016; Dautriche et al., 2017) mitigated this concern by only considering monomorphemic words, discarding a large fraction of the lexicon in the process.

We take a different approach to this problem by proposing a two-step model designed to mitigate the effect of morphemes. We begin by training an unregularized linear regression model to predict semantic vectors from morpheme-level features. Then, we use the residuals of this first stage morpheme-level model as the new target vectors for the sparsely regularized phonestheme extraction model. This removes the components of the word vector that are predictable from morpheme-level information, leaving only the aspects of word meaning not covered by morphology.

We use the the morphological analyses in the CELEX lexical database (Baayen et al., 1996) to compile a list of morphemes, which is used to create the morpheme-level feature vectors. We also use this list to remove any morphemes that may appear in the final model output.

3 Data

For our experiments, we use 300-dimensional GloVe (Pennington et al., 2014) English word embeddings trained on the cased Common Crawl. Many of the terms in the set of pretrained vectors are not English words. As a first attempt toward removing non-English words and named entities, we discard types that are not alphabetical or not completely lowercased. In addition, it's unlikely that rare words or very common words will contribute to the formation of sound-meaning associations (Hutchins, 1998). To further filter these rare or common words (and remove additional non-English types), we remove types that either occur less than 1000 times in the Gigaword corpus or in more than half of all Gigaword documents. Lastly, we remove types that share the same lemma if the lemma is also in the set of filtered word vectors. After this process, we are left with 7889 types out of the original 2.2 million.

We phonemicize our vectors by associating each word's vector to the word's ARPAbet symbol sequence, as provided in the CMU Pronouncing Dictionary (Carnegie Mellon University, 2014). If multiple types have the same ARPAbet symbol sequence (and are thus homophones), we discard them all. We also do not use types that are not in the CMU Pronouncing Dictionary. Phonemicizing the filtered set of vectors results in a set of 6633 vectors.

Note that our model can be applied using either orthographic or phonemicized vectors. Phonesthemes are an inherently phonetic phenomenon, which suggests that it is ideal to model the features at the phoneme level. However, using character-level features, in some cases, will be a reasonable approximation, especially since many of our extracted phonesthemes have a consistent orthographic representation. We release code for preprocessing data and training the models at http://nelsonliu.me/papers/phonesthemes/.

4 Experiments and Results

The candidate phonesthemes considered by the model are the word-initial phoneme bigram sequences that occur more than five times in our set of phonemicized vectors; we set a frequency threshold for feature inclusion since rare prefixes are unlikely to carry meaning. Each word's feature vector is a one-hot encoding of its bigram phoneme prefix. We choose to focus on word-initial bigrams since the bulk of prior work in linguistics has also focused on phonesthemes in this position. However, our method easily extends to larger subword units (e.g., trigrams), candidate phonesthemes within or at the end of a word, even other languages; we leave analysis of phonesthemes of other sizes, in different positions, and of different languages for future work.

We train our two-stage model on the phonemicized vectors; the features that are assigned a nonzero weight are our model-predicted phonesthemes. The features of our morpheme-level model are binary indicator features corresponding to 181 different morphemes extracted from the CELEX2 database. In total, our phonestheme extraction model considers 307 candidate phonesthemes; tuning the regularization strength on held-out error in 5-fold cross-validation results in a model that selects 123 candidate phonesthemes as predictive. The phoneme bigrams corresponding to the 30 features with the highest absolute model weight are in Table 1. Qualitatively, the words with the lowest error under the model containing each selected phonestheme candidate seem semantically coherent.

Many of the phonesthemes identified by our model have been proposed and validated by past work. 13 of the top 15 model-predicted phonesthemes were in Hutchins' set of 17 proposed word-initial phoneme bigram phonesthemes. This is an improvement over past work; Otis and Sagi (2008) identified 8 as statistically significant, with a hypothesis space restricted to 50 pre-specified word beginnings and endings. Gutiérrez et al. (2016) also identified 8, but with a much larger hypothesis space of 225 candidates. Our model considers an even larger hypothesis space of 307 candidate phonesthemes, which are all automatically extracted from the set of word vectors.

	ARPAbet Sequence	Character Sequence	Model Example Words
† *	S N	*sn-*	*sneaks*, *snubs*, *sniffs*
*	S K	*sc-/sk-*	*screwing*, *squelched*, *scurry*
*	K R	*cr-*	*crunched*, *cringed*, *crummy*
*	S P	*sp-*	*spiffy*, *splendidly*, *spunky*
	B R	*br-*	*brags*, *brouhaha*, *brutish*
*	G R	*gr-*	*griping*, *grumbles*, *grandly*
*	T R	*tr-*	*tryst*, *trounce*, *truism*
*	S T	*st-*	*stupendous*, *startlingly*, *stunner*
† *	B L	*bl-*	*blase*, *blithely*, *blankly*
*	F L	*fl-*	*flaunted*, *flowered*, *fluff*
† *	G L	*gl-*	*glossed*, *gleam*, *glamor*
*	S L	*sl-*	*slouch*, *slogged*, *slime*
† *	D R	*dr-*	*droll*, *dreamer*, *drifter*
† *	S W	*sw-*	*swoon*, *swoops*, *swipes*
	W IH1	*wi-*	*wimpy*, *willy*, *wince*
	K AE1	*ca-*	*candied*, *caffeinated*, *cataclysm*
	P AE1	*pa-*	*pantry*, *pathogen*, *pancake*
	S IH1	*sy-/si-*	*syllable*, *simulators*, *synchronize*
	F R	*fr-*	*froth*, *frock*, *freaks*
	M AE1	*ma-*	*mallet*, *masts*, *manor*
	P EH1	*pe-*	*pendant*, *pelt*, *petulant*
	M EH1	*me-*	*meld*, *meditate*, *memorized*
	M AH1	*mu-*	*mumbled*, *mummies*, *mutter*
*	K L	*cl-*	*clumsily*, *clunky*, *claustrophobic*
	S EH1	*se-/ce*	*sensuous*, *celibate*, *celebrants*
	AH0 B	*ob-*	*obliterate*, *abridged*, *obliquely*
	B AA1	*ba-/bo-*	*barbarous*, *bogs*, *barbers*
	P L	*pl-*	*pled*, *pliable*, *platoons*
	K AO1	*co-*	*corset*, *coroners*, *corduroy*
	F EH1	*fe-*	*fairest*, *fender*, *feds*

Table 1: The 30 model-predicted phonesthemes with the highest absolute model weight and their typical orthographic representation. The model example words were selected from the 10 phonestheme-bearing words with the lowest model error. ∗ indicates a phonestheme identified by Hutchins (1998). † indicates a phonesthemes with statistical support from Otis and Sagi (2008).

Validating Phonesthemes with Human Judgments Following the method of Hutchins (1998) and Gutiérrez et al. (2016), we empirically evaluate our phonesthemes by soliciting naïve human judgments about how well-suited a word's form is to its meaning.

We randomly selected 5 words containing each

of the top 15 model-selected phonesthemes and 5 words containing 15 random phonestheme candidates that were not selected by the model, for a total of 150 words.

We recruited native English-speaking participants through Mechanical Turk, and asked them to judge how well each word fits its meaning on a Likert scale from 1 to 5. 150 words is too many judgments for a single HIT (annotators would become fatigued and words might start to lose meaning). As a result, we randomly divided the task into 10 different HITs, each with 15 of the words to be tested. We required Amazon Mechanical Turk Masters status for the crowdworkers and compensated them $0.20 per HIT; each word received 30 ratings.

Following Hutchins (1998), we compute ratings for each candidate phonestheme by averaging the rating of the words that contain it. On average, model-predicted phonesthemes were rated 0.58 points higher than unselected phonestheme candidates (3.66 versus 3.08, respectively). To assess whether this difference is statistically significant, we use the one-tailed Mann-Whitney U test (Mann and Whitney, 1947) since the data is ordinal and unpaired. Based on the results of the test, we reject the null hypothesis that the average rating of words containing model-selected phonesthemes is *not greater* than the average rating of words that contain phonesthemes not selected by the model ($p < 10^{-9}$).

Figure 1 plots the human ratings of the top 15 model-selected phonesthemes against their absolute weight under the model; there is a weak positive correlation ($r = 0.081$).

2 of the 15 model-predicted phonesthemes with the highest absolute weight were not previously proposed by (Hutchins, 1998): *br-* and *wi-*. Both of these sound clusters seem like plausible phonesthemes. To the authors, the *br-* cluster evokes the idea of a raw, almost uncultured force, with words like "*brags*," "*brutish*," and "*brusque*" appearing among the words with the lowest error under the model. The types containing the word-initial *wi-* cluster with the lowest error under the model seem to convey fragility: "*wimpy*," "*wince*," and "*weak*."

From Figure 1, we can see that the *br-* phonestheme candidate received a very high model weight, but received lower ratings on average from human annotators. On the other hand, the average human rating of the *wi-* phonestheme candidate seems in line with its assigned model weight. Future work could further explore whether *br-* and *wi-* have psychological reality to native speakers.

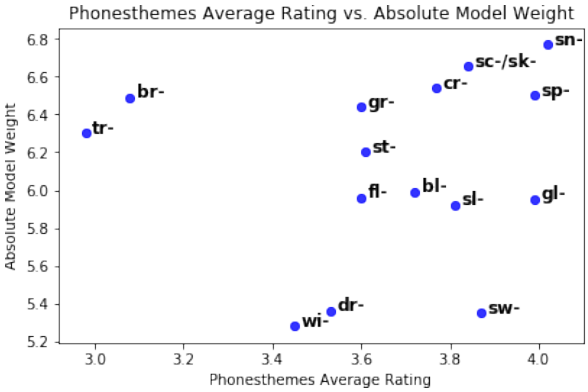

Figure 1: Average human rating versus the absolute model weight for the 15 selected phonesthemes with the highest absolute model weight.

5 Related Work

Several psycholinguistic studies have shown that native speakers associate certain sounds with a particular meaning, and phonesthemes have been identified in languages from English (Wallis, 1699; Firth, 1930) to Swedish (Abelin, 1999) and Japanese (Hamano, 1998). Bergen (2004) additionally demonstrates that phonesthemes affect online implicit language processing, and Parault and Schwanenflugel (2006) suggest that they play a role in language acquisition.

In recent years, the work of Otis and Sagi (2008) and Abramova and Fernández (2016) used computational methods to automatically detect and validate phonesthemes by examining whether words that contain a candidate phonestheme are more semantically similar than predicted by chance, according to a distributional semantic model. Dautriche et al. (2017) analyze lexicons of Dutch, English, German, and French and find that the space of monomorphemic word forms is clumpier than what would be expected by chance, according to lexical, phonological, and network measures.

Most similar to our work is that of Gutiérrez et al. (2016), who introduce an algorithm for learning weighted string edit distances that minimize kernel regression error and use it to detect systematic form-meaning relationships within language. Our model uses linear regression between candidate phonestheme features and semantic vectors. In addition, our model directly selects the

predicted phonesthemes with sparse regularization; their model instead provides a systematicity score for each type, and they extract phonesthemes by taking the word-beginnings with mean errors lower than predicted by a random distribution of errors across the lexicon.

6 Conclusion

In this work, we present a simple model for extracting non-systematic form-meaning relationships from a collection of word vectors. Our model is a sparsely regularized linear regression model that seeks to predict a word's semantic vector from a feature vector that encodes information about the candidate phonesthemes it contains; the sparse solutions of the regression problem have the effect of automatically selecting the features that are most predictive of word meaning, which we take as predicted phonesthemes.

We also develop a simple and effective two-stage approach for mitigating the effect of morphemes in the model. We initially train a model to map from morpheme-level features to word vectors, and then use the residuals of the morpheme-level model as the targets for the downstream phonestheme extraction model.

We empirically compare our model's predicted phonesthemes and find that many were previously proposed by linguists. We verified our results with human judgments of proposed and unselected phonesthemes, and annotators believe that words with a model-selected phonestheme "fit their meaning" more than words that contain a candidate phonestheme that was not selected by the model.

Acknowledgments

NL received support from a University of Washington Mary Gates Research Scholarship and a Washington Research Foundation Fellowship. This work was supported in part by NSF grant IIS-1562364. This work was also supported in part by a hardware gift from NVIDIA Corporation.

References

Åsa Abelin. 1999. *Studies in sound symbolism*. Ph.D. thesis, University of Gothenburg.

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: A data-driven study of conventional iconicity.

R. Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1996. CELEX2.

Benjamin K Bergen. 2004. The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

Carnegie Mellon University. 2014. The CMU Pronouncing Dictionary v0.7b. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Annette D'Onofrio. 2014. Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and Speech*, 57(3):367–393.

John Firth. 1930. *Speech*. Oxford University Press.

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

Shoko Hamano. 1998. *The Sound-Symbolic System of Japanese*. Cambridge University Press.

Charles F. Hockett. 1960. The Origin of Speech. In *Scientific American*, volume 203, pages 88–96.

Sharon Suzanne Hutchins. 1998. *The psychological reality, variability, and compositionality of English phonesthemes*. Ph.D. thesis, Emory University.

Wolfgang Köhler. 1929. *Gestalt psychology*.

Margaret Magnus. 2000. *What's in a Word? Evidence for Phonosemantics*. Ph.D. thesis, University of Trondheim.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Padraic Monaghan, Richard C Shillcock, Morten H Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1651).

John J Ohala. 1984. An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. *Phonetica*, 41(1):1–16.

Katya Otis and Eyal Sagi. 2008. Phonaesthemes: A corpus-based analysis. In *Proc. of CogSci*.

Susan J Parault and Paula J Schwanenflugel. 2006. Sound-symbolism: A piece in the puzzle of word learning. *Journal of psycholinguistic research*, 35(4):329.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proc. of EMNLP*.

Vilayanur S Ramachandran and Edward M Hubbard. 2001. Synaesthesia – A Window into Perception, Thought and Language. *Journal of Consciousness Studies*, 8(12):3–34.

Ferdinand de Saussure. 1916. *Course in General Linguistics*.

Richard Shillcock, Simon Kirby, Scott McDonald, and Chris Brew. 2001. Filled pauses and their status in the mental lexicon. In *Proc. of DiSS*.

John Wallis. 1699. *Grammar of the English Language*. Oxford.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.