Questioning Arbitrariness in Language: a Data-Driven Study of Conventional Iconicity

Ekaterina Abramova
Radboud University Nijmegen
e.abramova@ftr.ru.nl

Raquel Fernández
University of Amsterdam
raquel.fernandez@uva.nl

Abstract

This paper presents a data-driven investigation of *phonesthemes*, phonetic units said to carry meaning associations, thus challenging the traditionally assumed arbitrariness of language. Phonesthemes have received a substantial amount of attention within the cognitive science literature on sound iconicity, but nevertheless remain a controversial and understudied phenomenon. Here we employ NLP techniques to address two main questions: How can the existence of phonesthemes be tested at a large scale with quantitative methods? And how can the meaning arguably carried by a phonestheme be induced automatically from word embeddings? We develop novel methods to make progress on these fronts and compare our results to previous work, obtaining substantial improvements.

1 Introduction

It has long been held in linguistics that since the same concept can be expressed with words whose forms do not resemble each other (e.g., English *dog* vs. Italian *cane* vs. German *Hund*), there is no intrinsic link between how words sound and what they mean. This feature—*arbitrariness*—is often considered a hallmark of human language (Saussure, 1916; Hockett, 1959). At the same time, however, over the last decades, mounting evidence from psycholinguistic studies (Markel and Hamp, 1960; Ohala, 1984; Fordyce, 1989) has shown that speakers do in fact associate words that contain a particular form with certain meaning—that is, there is a degree of *iconicity* in language in addition to arbitrariness, which has been claimed to benefit language learning (Monaghan and Christiansen, 2006; Monaghan et al., 2014).

Non-arbitrary form-meaning associations come in two basic varieties: primary iconicity (also called 'true') and secondary (or 'conventional') iconicity. In the former, the sound is thought to directly resemble the meaning, as in onomatopoeia.[1] In the latter, the relationship is a statistical regularity according to which words that share similar sounds tend to be also similar in meaning, such as a large proportion of English words that end with the sound /-æʃ/ (e.g., *crash*, *slash*, *mash*, *trash*, *dash*) being related to destructive action or collision (Hutchins, 1998). The phonetic units that exhibit conventional meaning regularities of this latter kind are called *phonesthemes* and are the focus of the present paper. In particular, we investigate two main questions: (1) how can the existence of phonesthemes be tested at a large scale by means of a data-driven method? and (2) how can the meaning arguably conveyed by a phonestheme be derived automatically?

Phonesthemes are traditionally distinguished from morphemes in being non-compositional. That is, *unthinkable* can be thought of as being composed of morphemes *un-* (meaning *not*), *think*, and *-able* (meaning *capable of*), all contributing to the overall meaning of "incapable of being thought" and susceptible of being combined with other units with predictable semantic effects (e.g., *un-drinkable, think-er*). On the other hand, *crash* is not con-

[1] Although, the relationship can still be modified by phonetic features of a particular language, e.g., the rooster says *cock-a-doodle-doo* in English but *kikiriki* in German.

Proceedings of NAACL-HLT 2016, pages 343–352,
San Diego, California, June 12-17, 2016.

sidered to be formed compositionally from *cr-* and *-ash*, since these components do not possesses an easily identifiable independent meaning that can be combined productively with other morphemes.

Because phonesthemes challenge defining features of language such as arbitrariness and compositionality, they remain a rather controversial and poorly understood phenomenon. To a large extent, this is due to methodological issues. Early evidence for the existence of phonesthemes consisted primarily in linguists pointing out instances of intuitive correlations between a phonetic unit and the meaning of words containing that unit (Marchand, 1959; Reid, 1967), while early psycholinguistic experiments attempted to elicit meaning definitions for predefined lists of (real or nonsense) words sharing a phonetic unit traditionally considered to be a phonestheme (Fordyce, 1989; Abelin, 1999; Magnus, 2000). More systematic studies have subsequently been carried out by Hutchins (1998) and Bergen (2004), but overall the phenomenon of conventional linguistic iconicity as reflected in phonesthemes remains largely understudied, certainly within the computational linguistics community.

In this paper, we investigate phonesthemes by analyzing their orthographic correlates in a large corpus of written English, leveraging word embeddings constructed with `word2vec` and made available by Baroni et al. (2014). In particular, we make the following contributions:

- We develop a stricter test than previously done in the literature for deciding whether a unit exhibits conventional iconicity.
- We propose a new unsupervised method to induce the meaning conveyed by a phonesthemic unit.
- We evaluate our meaning induction method with new automatic evaluation techniques and compare its performance to a WordNet-based method proposed by Abramova et al. (2013), obtaining a very substantial improvement.
- We additionally evaluate our automatically derived meanings with human judgments collected via a crowdsourcing experiment.

We believe phonesthemes deserve thorough investigation for both theoretical and practical reasons. Theoretically, adding more data-driven methods can substantially enhance work in linguistics and psycholinguistics. Within computational linguistics itself, cross-fertilization with computational morphology (Wang et al., 2012; Marelli and Baroni, 2015), is an exciting avenue to be pursued. With respect to potential applications of automatic phonestheme meaning induction, creative brand naming (Klink, 2000; Özbal and Strapparava, 2012), sentiment analysis (Sokolova and Bobicev, 2009) and construction of more appropriate language teaching materials (Imai et al., 2008) are viable possibilities.

2 Related Work

Psycholinguistic studies on the nature of phonesthemes have shown that people tend to associate certain sounds with a particular meaning. Such studies were conducted on different languages, employing different methods and exhibiting various degrees of scale and systematicity (Fordyce, 1989; Abelin, 1999; Magnus, 2000; Hutchins, 1998). Recently, it has also been shown that phonesthemes affect online implicit language processing (Bergen, 2004) and language learning (Parault and Schwanenflugel, 2006).

What also emerged from these studies is that phonesthemes are not a homogeneous phenomenon. They can vary in terms of the number of words that contain a given phonestheme, their frequencies, the strength of their association with the core meaning of a phonestheme (for example measured as an average of human ratings for all the words that comprise the given phonesthemic cluster) and the regularity of that association (what proportion of words in the whole cluster are highly related to the predicted meaning). However, psycholinguistic data is to some extent ambiguous on how these features of phonesthemes affect their productivity, learnability and their effect on language processing, which could partly be due the methods being employed. For example, in order to determine the regularity of sound-meaning association, Bergen (2004) consults word definitions in Websters 7th collegiate dictionary and counts how many of those words bear the required meaning for a given phonestheme. The procedure requires an intuitive judgment from the experimenter in determining the meaning of a phonestheme and estimating whether a given word has that

meaning, and as a result is prone to experimenter bias and does not allow for large-scale testing. Finding a more automatic method for assessing phonestheme features and determining their meaning could thus alleviate this type of research liabilities.

Otis and Sagi (2008) and Abramova et al. (2013) are two studies that attempted to test for the existence of phonesthemes in a corpus-based automatic manner. For both the guiding question was: are words that contain a given phonetic unit, thought to be a phonestheme, more semantically similar than would be expected by chance? Using distributional models they compared the average cosine similarity of the vectors that correspond to phonestheme-bearing words to similarly sized groups of random words. Both studies found support for a sizable proportion of phonetic units tested. However, it could still be questioned whether the comparison was sufficiently strict, given that sets of random words which do not overlap in form have a priori lower chance of being semantically related than sets of words that share a phonestheme. Therefore, in the first part of our study (Section 4) we present a stricter validation method for candidate phonesthemes that also includes considerations related to morphological diversity, which were ignored in previous work.

Abramova et al. (2013) presented the first attempt to automatically assign meaning to sets of phonestheme-bearing words. The authors viewed the task as an instance of unsupervised ontology acquisition in the style of Widdows (2003) and used WordNet to assign over-arching labels to phonesthemic groups of words. While the approach was moderately successful in inducing WordNet labels that were in the direction predicted by the literature for a few phonesthemes (e.g., *gl*-containing words were assigned light-related labels), most phonesthemes did not receive meaningful labels according to the meanings typically associated with phonesthemes in the sound iconicity literature. The authors surmise that the failure could be due to the nature of WordNet, e.g., that it reflects only one type of semantic relation (hypernymy) which might not exhaust the links between words that share a phonestheme. In Section 5, we present a different approach to phonestheme meaning induction that exploits the properties of word embeddings in a fully unsupervised manner and yields substantially better results.

3 Data

Candidate phonesthemes. Following the studies of Hutchins (1998), we compile a list of possible phonesthemes of interest and their respective semantic glosses (more on the latter in Section 5). We will refer to these units as "candidate phonesthemes" because they all have been considered phonesthemes by previous qualitative studies and our aim is to investigate whether their alleged phenesthemic status is warranted quantitatively. Specifically, we focus on two-consonant units in word-initial position, which we will often call prefixes.[2] We focus on the 16 prefix candidate phonesthemes listed in Table 1. Since we work with orthographic correlates of phonetic units, we restrict ourselves to prefixes that have clear orthographic–phonetic mappings, discarding prefixes that allow for variation, such as *sc-/sk-*.[3]

bl-	*cl-*	*cr-*	*dr-*	*fl-*	*gl -*	*gr-*	*sl-*
sm-	*sn-*	*sp-*	*st-*	*sw-*	*tr-*	*tw-*	*wr-*

Table 1: Candidate phonesthemes considered.

Word embeddings. For our experiments, we use existing, high-quality word embeddings created and made available by Baroni et al. (2014).[4] We use the best performing model amongst those tested by Baroni and colleagues, which has been constructed with `word2vec`[5] using the CBOW approach proposed by Mikolov et al. (2013). The model contains 400-dimension vectors generated by considering the 300K most frequent word tokens (without lemmatization) in a large corpus comprising the English Wikipedia, the web-based corpus ukWaC (Baroni et al., 2009), and the BNC (Burnard, 2007).

Unfamiliar or very technical words are unlikely to contribute to the formation of sound-meaning associations (Hutchins, 1998). Accordingly, from the 300K target words present in the distributional model, we discard those that are not recognized by a

[2] Recall, however, that they do not correspond to morphological prefixes.

[3] Such alternation-susceptible prefixes are excluded from all analyses, including the baseline clusters introduced later on in Section 4.

[4] `http://clic.cimec.unitn.it/composes/semantic-vectors.html`

[5] `https://code.google.com/p/word2vec/`

comprehensive off-the-shelf English spell-checking dictionary. This results in a substantial reduction of the target vocabulary: 61,122 tokens remain after the filtering.[6] We use this restricted set of words and corresponding embeddings in all our experiments.

4 Phonestheme Validation

The aim of the first experiment is to investigate which of the candidate prefixes in Table 1 have phonesthemic character and thus evince conventional iconicity.

4.1 Methods

For a prefix to exhibit conventional iconicity, the words sharing that prefix must be semantically similar, while being morphologically diverse—i.e., their semantic relatedness must stem from their shared sound (as captured by the prefix's orthographic form), and not from their sharing of a common morpheme.

Semantic similarity factors. We start by assessing the degree of semantic similarity exhibited by all the words in the vocabulary that share a candidate phonestheme, which we refer to as candidate *phonesthemic clusters*. Our aim is to conduct a stricter test than previously done in the literature. Therefore, rather than comparing candidate phonesthemic clusters to sets of random words, as done by Otis and Sagi (2008) and Abramova et al. (2013), we compare them to words that share a random two-consonant prefix that is non-phonesthemic, i.e., not present in our list of candidate phonesthemes. Our vocabulary contains a total of 307 non-phonesthemic two-consonant prefixes. We refer to the sets of words sharing these prefixes as *baseline clusters*. For our subsequent analyses we use only 191 baseline clusters which contain between 10 and 2000 words. Naturally, such baseline clusters will contain words that are morphologically and hence semantically related, which offers a more challenging baseline.

[6] In particular, we use the spell-checking Python library PyEnchant for English; see `https://pythonhosted.org/pyenchant/api/enchant.html`. Many of the terms removed with this filtering mechanism correspond to non-words or named entities present in the corpus.

In our first similarity test, we compute cosine similarities for all possible pairs of words within every phonesthemic and baseline cluster. We then run 191 independent-samples one-tailed Welch's t-tests for each candidate phonestheme, comparing its pairwise similarity to the pair-wise similarity of each of the baseline clusters. For each candidate phonesthemic cluster, we record how many t-tests indicated significantly higher similarity than the baseline (using a Bonferroni-corrected threshold of $\alpha = .05/191$) as well as the effect size (Cohen's d) of the successful t-tests. Based on the binomial distribution (with α=.05), we obtain a significance threshold of 108—we hence judge a candidate prefix to exhibit significantly higher similarity than the baseline if more than 108 out of 191 t-tests are successful.

Our second similarity test is a check on the overall semantic cohesiveness of the candidate phonesthemic clusters. We calculate the average of all the pairwise similarities within our 191 baseline clusters. We then compare the average pairwise similarity of each candidate phonesthemic cluster to the distribution of the average similarity of the baseline clusters. We expect a positive correlation between the number of successful t-test per candidate phonestheme and their average similarity.

Morphological diversity factors. Since high semantic similarity could be due to the presence of a large proportion of morphologically related words rather than to a sound-meaning association, we want to balance similarity-based factors with considerations about the morphological diversity of the word clusters we investigate. In general, the larger the size of a cluster, the higher the chance for morphological diversity and the lower the chance for finding high semantic cohesiveness. Hence, we would expect a negative correlation between cluster size and semantic similarity.

In previous studies (Hutchins, 1998; Otis and Sagi, 2008) the impact of morphology is counteracted by manually removing morphologically related words before testing for semantic cohesiveness. Since one of our aims is to minimize manual intervention, we instead take into account morphological relatedness at the validation phase. To that end, we implement a crude lemmatization procedure and use the ratio between the number of words and

the number of lemmas in a cluster to estimate morphological diversity.[7]

The higher this ratio, the lower the morphological diversity—with the maximum value being equal to the size of the cluster when all words are reducible to a single lemma. We calculate this proxy of morphological diversity for candidate phonesthemic clusters and baseline clusters.

Validation constraints. Given the factors described above, we judge a candidate prefix to be a phonestheme if all the following conditions hold:

- *pairwise semantic similarity* is significantly higher than the baseline (according to our first semantic similarity analysis test)
- *average effect size* (Cohen's d) of pairwise similarity tests is at least 0.2
- *average semantic similarity* is higher than 2 standard errors above the average baseline similarity ($\mu = 0.1260, SE = 0.0038$)
- *ratio words/lemmas* is lower than 3 standard errors above the average ratio calculated for baseline clusters ($\mu = 2.93, SE = 0.0615$)

We have chosen each of the thresholds to be reasonably strong but not too restrictive since we rely on a combination of constraints. We deemed an average effect size of 0.2 sufficient given the strictness of our comparison method. The average semantic similarity of the phonesthemic cluster was required to be at least 2 standard errors above average similarity of the baseline clusters to approximate the conventional one-tailed alpha level of 0.025. Finally, a stricter threshold of 3 standard errors was chosen for the lemma ratio just to exclude cases of prefixes that have abnormally low morphological diversity. A stricter condition (requiring high diversity) does not seem justified since there is no reason to expect phonestheme-bearing words to be *more* morphologically diverse than average. The candidate phonestheme was judged to be significant when all constraints were simultaneously satisfied.

[7]Since our data consists of word embeddings (generalizing over contexts), an off-the-shelf lemmatizer is not effective. Instead we implement a lemmatizer dictionary based on the raw and lemmatized versions of the British National Corpus (Burnard, 2007). The lemma is retrieved if a given word is in the dictionary. Otherwise, we apply two state-of-the-art stemmers, first Lancaser, then Porter (this order was chosen after qualitative examination of a few examples).

4.2 Results

We apply the validation methods to our data. As a sanity check, we test whether the information encoded in the word embeddings is consistent with the data used in previous experiments: Indeed, the average similarity of the 16 candidate prefixes tested is positively correlated with the human ratings for semantic cohesiveness collected by Hutchins (1998) ($r = .46$), and with the similarity values reported by Otis and Sagi (2008) ($r = .68$) and Abramova et al. (2013) ($r = .58$).[8]

As predicted, the correlation between the average number of successful t-tests and average similarity is high ($r = .93$), suggesting that both methods are equally valid for evaluating semantic cohesiveness of phonesthemic clusters. Cluster size (both raw and as the number of lemmas) is negatively correlated with all semantic similarity measures, i.e. average pairwise similarity, average number of successful t-tests, and average effect size ($r \approx -.7$). This is consistent with the experimental finding by Hutchins (1998), who obtained lower human ratings for larger clusters of words.

Regarding evidence for conventional iconicity, the following six prefixes meet all our validation constraints: *bl-*, *gl-*, *sm-*, *sn-*, *sw-*, and *tw-*. Of the remaining 10 prefixes tested, 3 fail all constraints (*cl-*, *cr-*, *tr-*), 3 fail only the morphological diversity constraint (*gr-*, *sp-*, *st-*), and the rest fail some combination of constraints.[9] The 6 validated prefixes are a proper subset of the candidate phonesthemes validated according to the less strict methods used in earlier approaches: Otis and Sagi (2008) found support for *dr-* and *wr-* in addition to our 6 supported phonesthemes and Abramova et al. (2013) discarded only *cr-*, *sp-*, and *tr-* amongst our 16 candidates. This shows that our proposed validation procedure provides a compatible as well as stricter test for evidence of phonesthemic conventional iconicity.

5 Phonestheme Meaning Induction

The quantitative results presented so far show that, according to our validation constraints, some can-

[8]Recall that each of these studies uses a different corpus.

[9]Supplementary material including full details of the validation results for all candidate phonesthemes is available at `http://tinyurl.com/phonesthemes-naacl2016`.

bl-	to blow, swell, inflate; or to be round, swollen, or globular in shape	*bloat*, *blob*, *blow*
gl-	having to do with light or with vision; or something visually salient	*glow*, *glitter*, *glimmer*
sm-	a belittling, insulting, or pejorative term	*smirk*, *smother*, *smug*
sn-	related to the nose, or breathing; also snobbishness, inquisitiveness	*snout*, *sniff*, *sneeze*
sw-	to oscillate, undulate, or move rhythmically to and fro	*sway*, *swing*, *swoosh*
tw-	to turn, distort, entangle, or oscillate; or the result of such an action	*twist*, *twitch*, *tweak*

Table 2: Meaning glosses from Hutchins (1998, pp. 66–70) for the six validated phonesthemic prefixes, with example words.

didate phonesthemes do have phonesthemic character: they are present in words that are semantically similar while not being highly morphologically related. But what is the 'meaning' that these phonetic units convey? In this section, we aim at investigating whether the kind of meanings that have been informally proposed for these units in the sound iconicity literature can be derived using fully unsupervised methods.

In addition, we present ways for automatically evaluating the derived meanings, and finally conduct a human evaluation experiment via crowdsourcing.

5.1 Methods

Gold standard. We construct a set of gold standard meaning labels for each validated phonestheme by taking as a starting point the informal glosses provided by Hutchins (1998), who, in turn, compiled them by inspecting previous literature by Firth (1930), Marchand (1959), Wescott (1971) and others. The glosses for our validated phonesthemes are given in Table 2.[10] From each gloss, we extract the content words (ignoring words that bear the given phonestheme to avoid circular results) and manually discard words that play only an instrumental role in the definition. For example, in the following gloss for the phonestheme *sn-*, we keep the words in italics and discard the rest: "related to the *nose*, or *breathing*; or by metaphorical extension to snobbishness, *inquisitiveness*". Since the resulting lists of words are to some extent arbitrary (derived from intuitions of a single scholar), we extend them by manually adding synonyms of each of the initial seed words until each phonestheme is associated with 25 gold labels.[11]

Meaning induction. We generate an abstract meaning representation for a phonestheme by computing the centroid of the phonestheme-bearing word cluster. Our method for inducing the core meaning conveyed by a phonestheme is then very simple: We extract the nearest neighbors of the phonestheme centroid, with the constraint that these neighboring words cannot be members of the cluster themselves (i.e., must not exhibit the prefix in question). This method outputs an ordered set of words or *meaning labels*, which we can then evaluate against the gold standard labels.

As described in Section 2, the only previous attempt at automatically deriving the meaning of phonesthemes is due to Abramova et al. (2013). Their approach is inspired by the work of Widdows (2003) on ontology acquisition and it consists in assigning to a phonesthemic cluster the WordNet synsets that subsume as many as possible of the cluster words as closely as possible (i.e., within as few as possible intervening levels in the WordNet hierarchy). As discussed in that paper, this method does not only have the disadvantage of relying on a handcrafted ontology. Other shortcomings include WordNet's limited coverage in terms of vocabulary and type of semantic relations considered (mostly, hyponymy and synonymy).

For comparison purposes, we apply the WordNet meaning induction method of Abramova et al. (2013) and compare its performance to the unsupervised centroid method we propose.

[10] The semantic glosses for the remaining candidate phonesthemes can be found at `http://tinyurl.com/phonesthemes-naacl2016`.

[11] We consult an online edition of Roget's thesaurus (`thesaurus.com`) and retrieve only the synonyms that correspond to the relevant sense of a seed word, e.g., *light* in the sense of *illuminated*, not in the sense of *blond*.

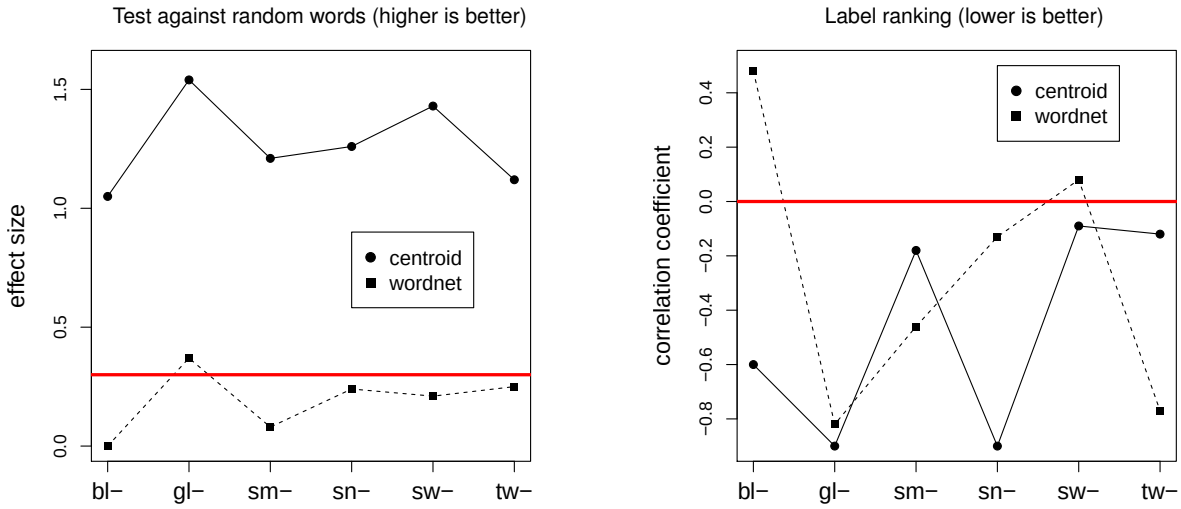

Figure 1: Evaluation of meaning induction methods. Left: only labels induced for prefixes above the horizontal line are significantly better than random labels. Right: negative coefficient scores shown below the line indicate better labels at the top of the list.

Automatic evaluation measures. Abramova et al. (2013) only offer an informal qualitative evaluation of their WordNet-based meaning labels. Here we propose two complementary ways of quantitatively evaluating induced phonesthemic meanings.

For our first meaning label evaluation test, we use a Monte Carlo analysis to determine whether generated labels are closer in vector space to gold labels than random sets of words. More specifically, for each phonesthemic cluster, we compute the *generated–gold similarities*, i.e., pairwise cosine similarities between all the generated labels and the gold set. We then create 100 sets of words, each composed of 25 words randomly drawn from the vocabulary and compute the *random–gold similarities*, i.e., pairwise cosine similarities between these sets of random words and the gold set. Next, we run 100 independent-samples one-tailed Welch's t-tests recording how many t-tests indicated significantly higher generated-gold similarity than random-gold (using a Bonferroni-corrected threshold of $\alpha = .05/100$). We also record the effect sizes (Cohen's d) of the successful t-tests. We repeat the procedure 3 times and take the average of these measures. Based on the binomial distribution (with $\alpha = .05$ and $p = .5$), we judge obtaining at least 59 successful t-tests to indicate that the generated labels are better than random baseline at capturing the phonesthemic meaning.

Our second evaluation test exploits the fact that both our centroid method and the WordNet method output ordered sets of labels (from more to less suitable). We are interested in testing whether the generated meaning labels are more similar to the gold labels the closer they are to the top of the list. To that end, we again compute the pairwise average similarity of each generated label with the gold label set and look at the correlation of that measure with the position k of the generated label. We expect similarity to decrease as k increases: Hence a strongly negative correlation indicates that the method retrieves the best labels first.

5.2 Results

Automatic analysis. We compare the meaning labels induced with our unsupervised centroid method to those generated with the WordNet method. An overview of the results can be seen in Figure 1. Regarding the first label evaluation test (*generated–gold similarities* vs. *random–gold similarities*; left plot in the figure), centroid overwhelmingly outperforms WordNet: Our method obtains significant results with high effect size for all phonestheme prefixes considered, while with the WordNet method only the labels derived for *gl-* are significantly more similar to the gold standard than random words.[12]

Regarding the order-sensitive evaluation measure

[12] In line with the informal evaluation by Abramova et al. (2013), who find that their meaning labels only seem to make sense for *gl-*.

(ranking of induced labels, right plot in Figure 1), with the centroid method we obtain negative correlations for all phonesthemes, although rather weak for *sw-* *tw-*, and *sm-*. This shows that the top induced labels tend to be closer to the gold meaning. Although the WordNet method again obtains results that are poorer overall, there are strong negative correlations for two phonesthemes: *gl-* and *tw-*.

Taken together, the results indicate that the WordNet method might be able to generate a few of good labels at the top of the list, especially when these labels are associated with the phonestheme-bearing words by hypernymy (e.g., the top *gl-* labels are *brightness, flash, radiance, lightness, look*). However, the remaining labels are mostly generic concepts such as *entity* and *object*, which do not produce significant results when compared as a group to the gold labels. The centroid method produces better labels overall as well as better labels at the top of the list. For example, the top *gl-* labels are *shimmered, twinkled, satiny* but it is also able to capture the meaning of other phonesthemes: the pejorative *sm-* receives *stunk* and *leered* as the top two and twisting and oscillating *tw-*'s top label is *waggled*.

Human evaluation. In addition to the automatic label evaluation procedures we have developed, we test our induced meaning labels against human judgments. What we aim at testing here is whether the semantic closeness to the gold standard meaning that we have been able to detect in vector space can actually be perceived by speakers.

We conducted a data collection study using the crowdsourcing platform CrowdFlower.[13] To prepare the stimuli, for each of the six validated phonesthemes we selected the 10 most frequent gold labels,[14] the 10 top labels induced with our centroid method, and 10 words randomly drawn from the vocabulary, with a BNC frequency of at least 100 to try to minimize the presence of words possibly unknown to the participants. The 100 cut-off is justified given that the average frequency of the gold labels is not significantly higher than the average frequency of all words above this threshold ($t = 1.876$, $p < 0.05$).

[13] http://www.crowdflower.com/

[14] According to the BNC frequency lists: https://www.kilgarriff.co.uk/bnc-readme.html

Prefix	AvgCount	Stats
bl-	5.53	$t(29) = 1.35$
gl-	6.57	$t(29) = 4.37$**
sm-	5.93	$t(29) = 2.04$*
sn-	6.7	$t(29) = 4.12$**
sw-	4.93	$t(29) = -0.18$
tw-	7.8	$t(29) = 5.66$**

Table 3: Results of human evaluation: avg. number of times an induced label was selected (N=30). *=p<0.05, **=p<0.001

An annotation item consisted of the set of gold words and 10 pairs of induced-vs-random labels (randomized in order). The participants were asked to judge which of the words in each pair was more related to the set of gold words.[15] We constructed 10 annotation items per phonestheme (including the same top 10 induced labels but paired with different random words) and for each annotation item we collected judgments from three different subjects (thus $N = 30$ items per phonestheme).

To analyze the results, we counted how many times an automatically induced label was selected as more similar to the gold label set than a random word. We performed a t-test with an alternative hypothesis that the mean number of selected induced labels per item is greater than 5 (i.e., greater than chance since there were 10 pairs to be judged per item). Automatically induced labels for 4 out of 6 phonesthemes (*gl-*, *sm-*, *sn-* and *tw-*) were judged to be related to the gold meaning to a higher degree than random words. Detailed results are in Table 3.

The fact that we obtain significant results indicates that our generated labels are meaningful not only according to automatic evaluation measures but also in terms of what speakers can perceive. However, the pattern of which phonesthemic labels receive better human judgments is somewhat less clear. For example, the appropriateness of the *gl-* labels is highly significant according to human judgment as well as both automatic tests (effect size and average similarity correlation with k). At the same time, while the *sw-* labels achieve a high effect size (see left plot in Figure 1), they are not judged signif-

[15] A screenshot of the instructions given to the participants can be found at http://tinyurl.com/phonesthemes-naacl2016.

icant in our human study. The pattern is reversed for *tw-*. Whether this exposes a real difference in sensitivity to phonesthemic meanings in human judgments compared to vector-based methods, remains an open question.

6 Conclusions

The analysis we have presented in this paper confirms that the connection between sound and meaning is not always entirely arbitrary and shows that this can be detected using the properties of word embeddings. We find, in line with previous computational and psycholinguistic studies, that words that share certain phonetic prefixes without being morphologically related are more semantically similar that would be expected by chance. In particular, our phonestheme validation procedure is stricter compared to previous work since we use sets of words that share a random two-consonant prefix as baseline and, importantly, take into account morphological relatedness. According to our more principled and stricter constraints, the following six consonant prefixes exhibit symptoms of conventional sound iconicity: *bl-*, *gl-*, *sm-*, *sn-*, *sw-*, and *tw-*. The validation method we employ could serve as a starting point for discovering new phonesthemes. For example, we could inquire whether any of the two-consonant clusters that we consider a baseline is in fact a previously unrecognized phonestheme.

The second aspect we have addressed concerns the automatic induction of the meaning conveyed by a phonestheme. Up to now, the arguable meanings of phonesthemes have been approximated informally by scholars (Hutchins, 1998; Bergen, 2004). To make progress on this front, we have proposed a fully unsupervised meaning induction method that relies on extracting semantic nearest neighbors of a phonesthemic cluster centroid in vector space. We have shown that this method achieves substantially better results than the WordNet-based method of our previous work (Abramova et al., 2013), generating meaning labels that are closer to the meanings proposed in the theoretical literature. For a subset of phonesthemes (4 out of 6: *gl-*, *sm-*, *sn-* and *tw-*), the higher suitability of the centroid-based meaning labels (as compared to random words) was also detected by human evaluators. Although there is obviously room for improvement, we think that these results are very promising given that this is the first data-driven study addressing this problem in an unsupervised manner.

References

Åsa Abelin. 1999. *Studies in sound symbolism*. Ph.D. thesis, Göteborg: Göteborg University.

Ekaterina Abramova, Raquel Fernández, and Federico Sangati. 2013. Automatic labeling of phonesthemic senses. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1696–1701. Cognitive Science Society.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL-2014*, volume 1, pages 238–247.

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

Lou Burnard, editor. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford Universtity Computing Services.

John R. Firth. 1930. *Speech*. Ernest Benn, London.

James Forrest Fordyce. 1989. *Studies in sound symbolism with special reference to English*. University Microfilm International.

Charles F. Hockett. 1959. Animal "languages" and human language. *Human Biology*, 31(1):32–39.

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

Mutsumi Imai, Sotaro Kita, Miho Nagumo, and Hiroyuki Okada. 2008. Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65.

Richard R. Klink. 2000. Creating brand names with meaning: The use of sound symbolism. *Marketing Letters*, 11:5–20.

Margaret Magnus. 2000. *What's in a word? Evidence for phonosemantics*. Ph.D. thesis, Trondheim, Norway: University of Trondheim.

Hans Marchand. 1959. Phonetic symbolism in english word formations. *Indogermanische Forschungen*, 64:146–168.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with

compositional distributional semantics. *Psychological Review*, in press.

Norman N. Markel and Eric P. Hamp. 1960. Connotative meanings of certain phoneme sequences. *Studies in Linguistics*, 15(1):47–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Padraic Monaghan and Morten H. Christiansen. 2006. Why form-meaning mappings are not entirely arbitrary in language. *Proceedings of the 28th annual conference of the Cognitive Science Society*, pages 1838–1843.

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 fo of voice1. *Phonetica*, 41:1–16.

Katya Otis and Eyal Sagi. 2008. Phonaesthemes: A corpus-based analysis. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 65–70.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 703–711, Stroudsburg, PA, USA. Association for Computational Linguistics.

Susan J Parault and Paula J Schwanenflugel. 2006. Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, 35(4):329–351.

David Reid. 1967. *Sound Symbolism*. T&A Constable.

Ferdinand de Saussure. 1916. Course in general linguistics.

Marina Sokolova and Victoria Bobicev. 2009. Classification of emotion words in russian and romanian languages. In *RANLP*, pages 416–420.

Hsueh-Cheng Wang, Li-Chuan Hsu, Yi-Min Tien, and Marc Pomplun. 2012. Estimating semantic transparency of constituents of english compounds and two-character chinese words using latent semantic analysis. In *Proceedings of CogSci*.

Roger Wescott. 1971. Linguistic iconism. *Language*, 47:416–428.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. HLT-NAACL*.