Automatic Labeling of Phonesthemic Senses

Ekaterina Abramova (e.abramova@ftr.ru.nl)
Department of Philosophy, Radboud University Nijmegen

Raquel Fernández (raquel.fernandez@uva.nl)
Institute for Logic, Language & Computation, University of Amsterdam

Federico Sangati (federico.sangati@gmail.com)
Institute for Logic, Language & Computation, University of Amsterdam

Abstract

This study attempts to advance corpus-based exploration of sound iconicity, i.e. the existence of a non-arbitrary relationship between forms and meanings in language. We examine a number of *phonesthemes*, phonetic groupings proposed to be meaningful in the literature, with the aim of developing ways to validate their existence and their semantic content. Our first experiment is a replication of Otis and Sagi (2008), who showed that sets of words containing phonesthemes are more semantically related to each other than sets of random words. We augment their results using the British National Corpus and the Semantic Vectors package for building a distributional semantic model. Our second experiment shows how the semantic content of at least some phonesthemes can be identified automatically using WordNet, thereby further reducing the room for intuitive judgments in this controversial field.

Keywords: Iconicity; Phonesthemes; Corpus analysis; Distributional Semantics; WordNet.

Introduction

The claim that the relationship between forms and meanings in language is not always arbitrary is controversial. However, evidence for non-arbitrary relationships comes at multiple levels of language, from phonology to syntax (Perniss, Thompson, & Vigliocco, 2010). Here we focus on the phonetic level and investigate the association of particular sounds with aspects of word meaning. Such sound iconicity has been described in a variety of non-Indo-European languages (see the studies in Hinton, Nichols and Ohala Hinton, Nichols, & Ohala, 2006b) and its existence in English suggested by a number of authors (Firth, 1930; Marchand, 1969), and exploited in commercial settings (Shrum & Lowrey, 2007).

Phonesthemes (a technical term for meaningful sound patterns) are sub-morphemic units that play a role of morphemes but have been traditionally distinguished from them by being non-compositional (but see Rhodes Rhodes, 2006 for an opposite view). The most oft-cited example is the English phonestheme *gl* which occurs in a large number of words related to light or vision (*glitter*, *glisten*, *glow*, *gleam*, *glare*, *glint*, etc.). Once the phonestheme is taken out, the remainder of the word is not a morpheme (*-itter*, *-isten*, *-ow* etc.) and one does not attach *gl* to other words to make them light-related. Still, the extent and the nature of this phenomenon is not clear.

Traditionally, the evidence for the existence of phonesthemes and their proposed meaning consisted in listing a number of words that share a given sound and attempting to find the semantic core that unites them. Popular explanations for the phenomenon would rest on the intuited association between sound production and meaning. For example, Reid (1967) states that "The explosive nature of the letter *b* is intensified when it is combined with *l* before the breath is released. Consequently words beginning with *bl* are found generally to indicate a 'bursting-out' or the resultant swelling or expansion" (p. 10). More recent accounts view them rather as a matter of statistical clustering. According to such "snowballing effect" theory, a group of phonemes in related words (for example, by common etymology) becomes over time associated with the meaning of these words and given the right conditions starts to attract other words with the same phoneme into a cluster, through semantic change or influencing the creation of new words (Blust, 2003; Hinton, Nichols, & Ohala, 2006a).[1]

Dissociating these competing explanations would require a combination of historical and cross-linguistic research but, arguably, there is a wealth of more basic questions that need to be addressed first. The nature of iconicity is such that it is easy to see the connection between form and meaning *once we are aware* of both elements but such intuition is not always a reliable guide for *discovering* the connections. Just as it is difficult to interpret an iconic sign from American Sign Language when its meaning is unknown (Bellugi & Klima, 1976), we might miss the connection that is in fact present. On the other hand, we might over-estimate the connection by listing only the light-related *gl*-words and forgetting the amount of *gl*-words that have nothing to do with light (*glide*, *glucose*, *globe*, *glove*, etc.). In other words, if we want to validate the existence of phonesthemes or explain their origin, we need to apply more falsifiable and unbiased methods in all stages of investigation: identifying them in a given language, quantifying their scope and establishing their meaning.

So far, the reality of phonesthemes has been demonstrated in behavioral experiments (Bergen, 2004; Hutchins, 1998) and corpus studies (Drellishak, 2006; Otis & Sagi, 2008). Our aim is to contribute to the second current of this research. We believe that this is a valuable way of objectively addressing large-scale linguistic phenomena that can refine our understanding of sound iconicity and lead to further testable hy-

[1] This is not to say that there are no universal sound features underlying certain cases of sound iconicity, such as words for small and large objects usually associated with high and low acoustic frequency respectively (Ohala, 1994).

potheses with respect to its cognitive underpinnings.

Otis and Sagi (2008) conducted the first corpus-based analysis of phonesthemes. They examined 47 groups of words containing phonesthemes using Project Gutenberg texts and a method for calculating word similarity based on Latent Semantic Analysis (LSA), in particular its Infomap[2] variety (Schütze, 1997). The analysis performed by Otis and Sagi showed that semantic relatedness of clusters of words that share a phonestheme is higher than that of clusters composed of randomly chosen words. This method, therefore, can be used to examine the validity of conjectured phonesthemes. However, as the authors admit, it "does not identify what specific semantic content is carried by the identified phonestheme" (p. 68). Our first aim is replicating the study of Otis and Sagi using (1) a more recent and balanced corpus – the British National Corpus (BNC), and (2) a newer and more versatile and efficient tool for calculating semantic relatedness, Semantic Vectors[3] (Widdows & Cohen, 2010).

Our second aim is attempting to develop a method for automatically identifying the semantic content associated with a particular phonestheme—a task that, to our knowledge, has not previously been addressed in the literature. Otis and Sagi (2008) suggest that methods designed to identify the topic of a given text could be used to that end. We think, however, that a more straightforward method lies in analogy with the task of unsupervised ontology acquisition: placing a word within a hierarchy of concepts based on its semantic relationship with the rest of the words in the hierarchy: for example, *pear* being placed close to *apple* and *banana* under *fruit*. In the case of phonesthemes, it is conceivable that a group of *gl* words would be assigned a vision-related higher class. Whether this can be done automatically and applied to a variety of phonesthemes is one of the questions we pose in this study.

In sum, our hypotheses are the following:

Hypothesis 1: Words that share a phonestheme are on average more semantically related than random words.

Hypothesis 2: The core semantic import conjectured in the literature for a phonestheme can be derived automatically from a set of phonestheme-bearing words.

Experiment 1: Semantic Relatedness

Methods

To explore our Hypothesis 1, we used the British National Corpus (Burnard, 2007), a 100 million word collection of written and spoken English language compiled from a wide variety of sources and genres. We pre-processed the entire corpus using the Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009). In particular, we used NLTK to extract the content words in the corpus (nouns, adjectives, and lexical verbs) and to lemmatize them, i.e. to reduce a family of inflected words (such as *walk*, *walks*, *walked*, *walking*) to a single word type or lemma (e.g. *walk*). This resulted in a subcorpus of about 43 million words, which we used as input to construct a distributional semantic model with the Semantic Vectors package (Widdows & Cohen, 2010).

Semantic Vectors allows us to use a corpus to build a high-dimensional vector space where words are represented as vectors that record their frequency of co-occurrence with other words or other documents in the corpus. We can then use well-defined methods to measure how similar the meanings of two words are, such as computing the cosine of the angle formed by their corresponding vectors. As Otis and Sagi (2008) indicated in their pioneering corpus study, this methodology can be of great value to investigate the claims behind phonesthemes in an objective, data-driven way, since we can use the distributional model to test whether words sharing a hypothesized phonestheme exhibit higher semantic similarity than random words.

Like Otis and Sagi, we built a *term-term* model where each word vector records the co-occurrence of that word with other words in the context (rather than recording occurrence in particular documents like LSA), but unlike them, who used the traditional *singular value decomposition* method for reducing the dimensions in the matrix, we used *Random Projection*, a more computationally efficient algorithm.[4] We experimented with the settings of two parameters in the Semantic Vectors package: the *minimum frequency* of the word types considered for building the model (as we may not be able to construct reliable distributional semantic representations for low frequency words) and the *window size*, i.e. the context window of n words to left and right of the target word where the model looks for co-occurrences of other words. McDonald and Ramscar (2001) claim that "the best fit to psychological data is typically achieved with word vectors constructed using context window sizes between ± 2 and ± 10 words." Otis and Sagi used $n = 15$, which is the default setting in Infomap.

We focused on the 22 prefix phonesthemes conjectured by Hutchins (1998). Our statistical analysis followed the procedure proposed by Otis and Sagi (2008). For each phonestheme, we first extracted all the vectors of the phonestheme-bearing word types in our distributional semantic model.[5] We shall refer to the resulting set of words (and vectors) as a phonestheme cluster. We then performed two Monte Carlo analyses. In the first analysis, we computed the average semantic similarity of each phonestheme cluster by forming 1000 random pairs and averaging the semantic distance obtained. In addition, we did the same for similarly-sized clusters of random words and performed an independent samples t-test for the resulting two groups of values. In the second analysis we took 50 random pairs within each phonestheme cluster and a corresponding group of pairs of random words and run 100 independent sample t-tests noting whether the mean of phonestheme cluster distances was significantly higher than the distances obtained for pairs of random words (with $\alpha = 0.05$). Based on the binomial distribution, we

[2]Freely available at http://infomap-nlp.sourceforge.net.

[3]Freely available at http://code.google.com/p/semanticvectors/.

[4]See Sahlgren (2005) and Widdows and Cohen (2010) for a comparison of these methods.

[5]Since we are dealing with a written corpus, this is done on the basis of an orthographic match with the phonetic grouping.

judged the number of significant t-tests as higher than 15 to lend statistical support to our Hypothesis 1. We performed the procedure 5 times and took the mean to be the final result.

We used the results obtained with the *gl* phonestheme cluster (which obtained the highest statistical support in the Otis and Sagi study) to optimize the minimum frequency and the window size parameters of our distributional semantic model. The model produced the most qualitatively sensible and most statistically stable results when setting the minimum frequency to 100 and the window size to 10. This resulted in a model containing a set of 22292 vectors. This vector space was used in all subsequent parts of our study.

Results

The results obtained with our parameter optimizing test on the *gl* phonestheme showed that the semantic relatedness of the words in the *gl* cluster was significantly higher than that of clusters of random words, as measured by our t-test procedure. On average, 26.4 t-tests produced a significant result (recall that the threshold of significance of the binomial test was for at least 15 out 100 t-tests to turn out as significant).

We used the same vector space (with the parameters fixed) to analyze the remaining 21 prefix phonesthemes. The results obtained are reported in Table 1. For each phonestheme, the table shows the number of word types in the phonestheme cluster (# Tokens), the average degree of semantic relatedness amongst those words (Sim) calculated according to our first Monte Carlo analysis, the number of significant t-tests (# Sig) calculated according to our second Monte Carlo analysis, and the mean effect size (Effect) of these t-tests. As can be seen, the model did not only confirm the semantic similarity of the words in the *gl* phonestheme (for which it had been optimized), but produced significant results for 16 different prefix phonesthemes out of the 22 considered. The average semantic relatedness of phonestheme clusters ($M = 0.446, SD = 0.044$) was highly correlated with the number of significant t-tests ($r = 0.95, p < 0.0001$) and was furthermore significantly higher than the average semantic relatedness of random words clusters ($M = 0.397, SD = 0.018$), as shown by an independent samples t-test ($t(42) = 4.83, p < 0.0001$).

Table 1: Phonestheme semantic relatedness results

Prefix	#Tokens	Sim	#Sig	Effect
bl-	105	0.4607	56.8	0.2845
cl-	156	0.4295	29.4	0.2570
cr-	197	0.3921	7.40	0.2327
dr-	99	0.4504	63.6	0.2849
fl-	137	0.4340	34.2	0.2536
gr-	197	0.4050	25.2	0.2617
sc-/sk-	167	0.4031	10.2	0.2443
scr-	32	0.5174	68.4	0.3093
sl-	83	0.4275	42.4	0.2734
sm-	42	0.4803	51.4	0.2817
sn-	40	0.4650	52.2	0.2909
sp-	161	0.4127	14.4	0.2392
spl-	11	0.5224	59.0	0.2723
spr-	24	0.3950	4.80	0.2373
squ-	24	0.4916	67.0	0.3205
st-	298	0.4307	25.0	0.2465
str-	89	0.4525	61.8	0.2899
sw-	67	0.5138	91.2	0.3396
tr-	249	0.3912	5.80	0.2318
tw-	22	0.4304	17.4	0.2408
wr-	38	0.5155	90.6	0.3915

In line with the findings of Otis and Sagi (2008), we were thus able to obtain support for Hypothesis 1 for 16 conjectured phonestheme prefixes. Using the BNC – a more general, balanced, and modern corpus of English than Project Gutenberg – our study yielded higher support for the hypothesis than Otis and Sagi's previous study, which had found only 12 phonestheme prefixes as reaching statistical significance.

Experiment 2: Phonestheme Cluster Labeling

Methods

After establishing significant differences between the semantic relatedness scores for phonestheme word clusters and clusters of random words, we turned to our second experiment, whose aim is to investigate possible methods to automatically detect the core semantic content carried by a phonestheme. To our knowledge, this is the first attempt to address this issue by objective means. To test our Hypothesis 2, we selected a number of prefix phonesthemes based on the amount of statistical support obtained in our first experiment and on how unambiguous and generally agreed upon were the sense definitions proposed in the literature. We selected 10 phonesthemes with high semantic relatedness scores and compiled a list of definitions based on the descriptions given by Hutchins (1998), Marchand (1969) and Reid (1967). The resulting list of phonesthemes together with their conjectured semantic import is presented in Table 2.

Table 2: Phonesthemic senses

Prefix	Definition	Example
bl-	swelling, explosion, extension, broadness	*bloating*
gl-	light, vision, look, brightness, shine	*glitter*
gr-	threatening noise, anger, grip	*growl*
scr-	unpleasant sound, irregular movement	*screech*
sn-	nose, mouth, smell, snobbish person	*sneeze*
spl-	divergence, spread, splash	*splash*
squ-	discordant sound, softness, compression	*squeeze*
str-	linear, forceful action, effort	*strike*
sw-	rhythmical movement	*swing*
wr-	irregular motion, twist	*wring*

In order to automatically assign a semantic class label to a phonestheme cluster, we used WordNet (Fellbaum, 2005), a cognitively motivated ontology of words and concepts linked by different semantic relations commonly used in computational linguistics. The main semantic relation connecting words that express different concepts in WordNet is the super/subordinate relation (also called hypernymy/hyponymy), which establishes a hierarchy of concepts from more general concepts like *animal* to increasingly specific ones like *mammal* or *whale*. Since hypernymy is a transitive relation, for

each word we can construct its hypernymy chain: the set of all its superordinate concepts or hypernyms connecting the word in question to the root node in the hierarchy (*entity* in the case of WordNet), ordered by their level of specificity.

WordNet is made up of independent hierarchies for different parts of speech: nouns, verbs, and adjectives. Given this (which prevents the possibility of assigning a cross-categorical semantic label) and the fact that the hierarchies for verbs and adjectives are far less complete than the noun hierarchy, we focused on the common nouns[6] within each phonesthemic cluster amongst those listed in Table 2. This resulted in eliminating 37% of words over all clusters.

For each common noun w in a phonesthemic cluster, we computed a set $H(w)$ containing all superordinate concepts in the hypernym chain of w, and derived a set $\mathcal{H}$ of potential class labels for that cluster by taking the union of all sets $H(w)$ for each noun w in the cluster. We then considered several methods for selecting the most optimal semantic class labels from $\mathcal{H}$. Our methods were inspired by the approach to unsupervised ontology acquisition proposed by Widdows (2003) according to which "the most appropriate class-label for the set [of words] S is the hypernym $h \in \mathcal{H}$ which subsumes as many as possible of the members of S as closely as possible in the hierarchy" (p. 278). Widdows offers a general scheme for defining an *affinity score function* $\alpha(w,h)$ between a word w and a candidate label h, which generates a ranking of all the potential class labels for a cluster of words:

$$\alpha(w,h) = \begin{cases} f(dist(w,h)) & \text{if } h \in H(w) \\ -g(w,h) & \text{if } h \notin H(w) \end{cases}$$

where dist(w,h) is a distance measure between a given word and a hypernym, f is a reward function that gives points to h if it subsumes w and the more points the closer this relationship, while g is a penalty function that subtracts points if h does not subsume w. The best class-label is the hypernym $h \in \mathcal{H}$ that has the highest affinity score summed over all the elements in the cluster.

Following Widdows, we chose as our distance measure the number of intervening levels in the WordNet hierarchy and set the rewarding function to $f = 1/dist(w,h)^2$. As for the penalty function g, we tested constant values of 0.25, 0.1 and 0.01. This particular variant of the scoring function thus magnifies the credit given to classes that are close to the words they subsume while giving a very small penalty to potential labels that miss out words in the cluster. The expected result is thus a ranking of class labels with a strong preference for specificity. This seems congruent with the nature of phonestheme clusters, which may contain a relatively large number of words that due to, for example, etymological factors are unlikely to be all related to the phonesthemic meaning. In fact, it is acknowledged in the literature that the sound-meaning associations are likely to be probabilistic (Hutchins, 1998) and that phonesthemic meaning can fall into related but separate groups. For example, *gr* is taken to be related to both angry noises (*growl*, *grunt*) and grabbing actions (*grab*, *grasp*). Given this, we also considered an approach whereby we first run a Gaussian Expectation-Maximization clustering algorithm on each phonestheme cluster to obtain more refined subsets of words and then run our scoring function algorithm on each of the resulting sub-clusters.

Finally, to counterbalance the preference for high specificity but potentially low coverage of the words in the phonesthemic clusters, we experimented with a different labeling algorithm that fixed a minimum coverage threshold. The algorithm examines all hypernyms $h \in \mathcal{H}$, selects those that subsume a minimum percentage θ of words in the cluster and then ranks them according to their specificity (the number of intervening levels to the root node *entity*). We tested the percentage values $\theta = 10$ and $\theta = 20$ and run the algorithm both on complete cluster phonesthemes and on the unsupervisedly derived sub-clusters.[7]

Results

Our results show that successful labeling of phonesthemic clusters can be performed but success depends on a number of factors. First, it is necessary to clarify what we mean by successful labeling. A labeling outcome of phonesthemic senses was deemed successful when the top 10 labels fulfilled the following criteria:

1. the topmost label is not the WordNet root node (*entity*);
2. the top 5 labels do not all have specificity score $m \leq 2$;
3. at least 50% of the top 10 labels carry meaning predicted for a given phonestheme;
4. the top 10 labels together subsume at least 50% of the words in the cluster.

These heuristics mean that if it is possible to establish the semantic core of a phonesthemic cluster using WordNet hypernym trees, the top labels will be both specific and in the direction predicted by the literature. It is always possible to subsume all the words in the cluster, independently of their semantic relationship, under the root, just due to the WordNet structure. Such a label, however, would not be very informative. By the same reasoning, we excluded the next two levels of the hierarchy which contain concepts such as *physical entity*, *abstraction*, *matter* or *relation*. On the other hand, specificity needs to be balanced out by coverage, i.e. it is possible to have very specific labels as top results but covering only a small portion of words in the cluster. Finally, the labels need to at least intuitively relate to the domain specified in the literature for a given phonestheme.

Given these criteria, we obtained clear positive results for one phonestheme (*gl*) out of 10 examined; moderately successful results for two phonesthemes (*sn* and *str*); and negative results for the remaining 7 phonesthemes. We present

[6] We discarded proper nouns, which in WordNet are always terminal leaf nodes representing concrete *instances* rather than types.

[7] Assigning a higher value to g would also increase coverage. However, for consistency g would have to be dependent on the size of the cluster. We instead choose a simple approach here which resorts to a percentage.

Table 3: Top 5 WordNet labels for *gl-*, *sn-*, and *str-*

Prefix	Label	Score	Spec	Cov
gl- (*N*=56)	brightness	4.82	6	23.7%
	flash	4.67	5	13.2%
	radiance	3.92	7	13.1%
	light	2.23	5	26.3%
	look	2.16	8	10.5%
sn (*N*=34)	laugh	1.71	5	6.5%
	unpleasant person	1.71	8	6.5%
	photograph	1.71	7	6.5%
	smell	1.71	8	6.5%
	piece	1.71	4	6.5%
	noise	1.71	6	6.5%
str (*N*=76)	effort	2.22	8	11.3%
	motion	1.14	7	11.3%
	labor	0.39	7	11.3%
	change	0.98	6	20.5%

the top labels for the 3 successfully labeled phonesthemes in Table 3, together with the scores calculated by the affinity score function with penalty set to a constant $g = 0.01$ (Score), specificity of each label (Spec) and the proportion of words in the cluster subsumed (Cov).

The *gl* phonestheme received light- and vision-related labels in all labeling algorithms that we tested. As can be seen in Table 3, they are clearly specific, cover a large proportion of words and all carry the predicted meaning. Similar results are obtained using other two settings of the g function and the coverage-based algorithm, although a small percentage of high-level labels like *entity* does appear in these lists.

Our moderately supported phonesthemes *sn* and *str* obtain labels in the predicted direction only with algorithms that reward specificity over coverage. For *sn* (related to nose, mouth and snobbism according to the literature), the best result is obtained with the distance measure and penalty function $g = 0.01$, while for the *str* phonestheme, related to forceful action – with the coverage-based algorithm of $\theta = 10$.

Similar conclusions can be drawn from phonesthemes that did not lead to clear tendencies in their labels or to specific enough labeling. Such lack of success is evident in either only one label out of the top 10 being relevant to the predicted meaning or all of the labels being very general. In the first case, for example, both *gr* and *scr* words are subsumed under *noise*. In fact, this label appears in all instances of *scr* scores as a top label, covering 26.9% of words in the cluster. However, the rest of the top labels are either of a general kind (*entity*, *change*) or not related to sound or movement (*handwriting*, *wound*) and therefore we cannot consider the labeling result to be very strong. In other cases, the words are primarily subsumed by labels like *entity* and *abstraction*.

As explained in the Methods, we considered the possibility that clusters might be composed of several groups of words that do not all share the same semantic content. This is especially likely for numerous clusters (e.g. *gr* cluster even with proper nouns removed contained 158 words). To counteract this problem we examined how prior sub-clustering affects the labeling results. The EM algorithm we used detected the presence of two clusters for 4 out of 10 phonesthemes we considered (*bl*, *gl*, *gr* and *str*).[8] Again, however, the most interesting result was obtained for the *gl* phonestheme. According to the labels we obtained for its two sub-clusters, only one of them was light-related. The second sub-cluster contained words like *gluten*, *glucose* and *glycoprotein*, which were placed under labels such as *protein*, *macromolecule* and *organic compound*, indicating a clear presence of chemistry-related words. No clear patterns were obtained for the other 3 phonesthemes with two sub-clusters. We do not exclude the possibility that this result was due to the quality of the vectors given to our clustering algorithm or to the algorithm itself. For example, the presence of a large proportion of kinship concepts in one of the *gr* sub-clusters (*grandfather*, *grandchildren* etc.) led to it being assigned labels such as *grandparent* and *ancestor*, while the same sub-cluster contained words such as *growl* and *grunt*. Therefore, whether the sub-clustering step is theoretically sound and if so how it should be accomplished requires further study.

Discussion

The results of our first experiment are largely in line with those of Otis and Sagi (2008), but we also see a number of differences. On the one hand, we obtain higher support for phonestheme clusters overall and show statistical significance for several phonesthemes previously unsupported. On the other hand, our support for the strongest phonesthemes in the original study (*gl* and *spr*) is weaker. These differences can be due to several factors. First, we use a different, more modern and balanced, BNC corpus and our resulting phonestheme clusters are larger. Second, we use a different method for building our distributional model – both a different algorithm (Random Projection) and a smaller window size.

It is worth noting that our tuning experiments with the *gl* phonestheme show that the kind of pre-processing that we apply to the corpus and the window size parameter do make a difference to the statistical results that can be obtained from the model. However, while pre-processing could be viewed as merely a methodological challenge common to all types of corpus analyses, there might be a theoretical significance behind the impact of the window size. Sahlgren (2008), for example, suggests that a small window size is preferable for detecting paradigmatic relationships between words (those that hold between words that do not co-occur themselves but occur in similar contexts, e.g. *dog* and *cat*) and at the same time there is evidence (Peirsman, Heylen, & Geeraerts, 2008) that larger context is beneficial for picking out syntagmatic relationships that hold between words that often occur together (e.g. "crystal clear"). To our knowledge, the kinds of relationships that hold between phonesthemic words (in general or depending on a given phonestheme) have not been systematically investigated using such distinctions and further work on the influence of the kind of context useful for detecting phonesthemic relatedness, in conjunction with experimental

[8] For the other phonesthemes, no sub-clusters were detected.

work on similarity, could offer clues on this issue.

Our labeling results are somewhat disappointing but, given the novelty of our approach, still highly informative. The fact that we obtain better results for *sn* and *str* phonesthemes with algorithms that favor specificity over coverage and that labeling is not fully successful with the remaining phonesthemes are puzzling given the high support that we obtain for these phonesthemes in our first experiment. We believe that there are two possibilities that can explain this.

The first possibility is that our WordNet-based methodology is not fully suited to discover the common semantic content that is present. WordNet does not allow for integrating hypernymy tree chains across different parts of speech, which might be vital for phonesthemes, a large proportion of which are verbs. In addition, it does not make all the distinctions that would be useful for phonesthemic studies, e.g. both *scr* and *gr* are associated with kinds of sound but one is "unpleasant" and the other one "threatening" – a distinction which is not part of the WordNet taxonomy. On a more general note, hypernymy might not be the most appropriate relation for all phonesthemes, e.g. *snout* and *sneezing* are not similar because they are both a type of nose. Therefore, perhaps better labeling results could be achieved using a different semantic network, such as ConceptNet[9], which allows for exploiting other than merely "is a" relations.

The second possibility, which we cannot reject, is that there is in fact no semantic core that unites phonestheme clusters and that the statistical support obtained by Otis and Sagi and in our first experiment is a result of a particular methodology. This interpretation is suggested by the fact that in our second experiment the overall coverage of phonesthemic words by the semantic labels is relatively low. Qualitative examination of the clusters also seems to show that they contain a lot of variability. In the future, we plan to design stricter tests – for example, comparing phonestheme clusters to clusters that share a particular (non-phonesthemic) sub-string rather than simply to a group of random words.

Ultimately, the aim of automatically detecting phonesthemes *and* their semantic content in a more objective, falsifiable way is, on the one hand, to help researchers interested in iconicity to validate the existence of phonesthemes previously reported in the literature, to possibly discover new phonesthemes, and to settle disputes over their particular meaning; and on the other hand, to open the door to investigating further the cognitive nature of the semantic relationships that unite phonestheme clusters. This study constitutes a step in this research programme.

References

Bellugi, U., & Klima, E. (1976). Two faces of sign: Iconic and abstract. *Annals of the New York Academy of Sciences*, *280*, 514-538.

Bergen, B. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 291-311.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Blust, R. A. (2003). The phonestheme n-in austronesian languages. *Oceanic Linguistics*, *42(1)*, 187-212.

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

Drellishak, S. (2006). *Statistical techniques for detecting and validating phonesthemes*. Unpublished master's thesis, University of Washington.

Fellbaum, C. (2005). Wordnet and wordnets. In *Encyclopedia of language and linguistics* (Second ed.). Oxford: Elsevier.

Firth, J. (1930). *Speech*. London: Oxford University Press.

Hinton, L., Nichols, J., & Ohala, J. J. (2006a). Introduction. In *Sound symbolism.* Cambridge University Press.

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (2006b). *Sound symbolism*. Cambridge University Press.

Hutchins, S. S. (1998). *The psychological reality, variability, and compositionality of english phonesthemes*. Unpublished doctoral dissertation, Emory University.

Marchand, H. (1969). *The categories and types of present-day english word-formation*. Munich: C. H. Beck.

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. 23rd CogSci.*

Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In *Sound symbolism.* CUP.

Otis, K., & Sagi, E. (2008). Phonaesthemes: A corpus-based analysis. In *Proc. 30th CogSci.*

Peirsman, Y., Heylen, K., & Geeraerts, D. (2008). Size matters: Tight and loose context definitions in English word space models. In *Proc. ESSLLI Workshop on Distributional Lexical Semantics.*

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, 1-15.

Reid, R. (1967). *Sound symbolism*. T&A Constable.

Rhodes, R. (2006). Aural images. In *Sound symbolism.* CUP.

Sahlgren, M. (2005). An introduction to Random Indexing. In *Proc. Methods & Applications of Semantic Indexing.*

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20(1)*, 33-54.

Schütze, J. (1997). *Ambiguity resolution in language learning*. Stanford: CSLI Publications.

Shrum, L., & Lowrey, T. (2007). Sounds convey meaning: The implications of phonetic symbolism for brand name construction. In *Psycholinguistic phenomena in marketing communications.* Mahwah: Erlbaum.

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

Widdows, D., & Cohen, T. (2010). The semantic vectors package: New algorithms and public tools for distributional semantics. In *Proc. 4th Int'l Conf. on Semantic Computing.*

[9]Freely available from `http://conceptnet5.media.mit.edu`.