Phonaesthemes: A Corpus-Based Analysis

Katya Otis (kotis@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Eyal Sagi (ermon@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Abstract

The association between sound and meaning is commonly thought of as symbolic and arbitrary. While this appears to be mostly correct, there is some evidence that specific phonetic groupings can be indicative of word meaning. In this paper we present a corpus-based method that can be used to test whether such an association exists in a given corpus for a specified phonetic grouping. The results we obtain using this method are compared with other empirical findings in the field and its implications are discussed.

Keywords: Corpus analysis, Computational linguistics, Phonaesthemes, Phonetics, Psycholinguistics, Sound-Meaning association.

It is a popular intuition that words with similar sounds also mean similar things. There is a long tradition of belief in the association between phonetic clusters and semantic clusters going back at least as far as Wallis' grammar of English (Wallis, 1699). Morphemes form one such well-known cluster, but other sub-morphemic phonetic clusters that contribute to the meaning of the word as a whole have also been hypothesized (Firth, 1930; Jakobsen & Waugh, 1979). Anthropologists have documented sound symbolism in many languages (Blust, 2003; Nuckolls, 1999; Ramachandran & Hubbard, 2001), but its role as a purely linguistic phenomenon is still unclear. Moreover, the Saussurean notion of the arbitrary relationship between the sign's form and its referent is a matter of dogma for most linguists (Hockett, 1960). This makes the study of words that *do* participate in predictable sound-meaning mappings all the more important, since, under the framework of contemporary linguistics it is difficult to explain how these patterns come to be, or why they might survive despite the obvious benefits of arbitrary sound-meaning mappings. What we mean by "sound-meaning mapping" is not purely sound symbolism, however, nor is it morphology. In the following paper, we offer a statistical, corpus-based approach to the *phonaestheme*, a sub-morphemic unit that has a predictable effect on the meaning of a word as a whole. These non-morphological relationships between sound and meaning have not been thoroughly explored by behavioral or computational research, with some notable exceptions (e.g. Hutchins, 1998; Bergen, 2004).

By contrast, sound-syntax mappings are somewhat better documented in the literature. Monaghan, Chater, and Christiansen (2005) address the role of phonetic similarity in separating lexical categories, a construct that necessarily includes some syntactic features and some semantic features (Monaghan et al., 2005).[1] Recent research indicates that systematic sound-meaning and sound-syntax relationships play a role in language processing (Hutchins, 1998; Bergen, 2004; Farmer, Christiansen, & Monaghan, 2006), and may also be important to language learning (Monaghan et al., 2005).

To the degree that it differs from adult-directed speech, child-directed speech should be sensitive to the child's status as a language learner. Monaghan et al. (2005) tested adult speech from the CHILDES corpus for the presence of 16 phonological cues in open- and closed-class words and for their diagnosticity in determining whether a word is a noun or a verb. Significantly diagnostic cues to the noun/verb distinction were: syllable length, onset and syllabic complexity, syllable reduction, *-ed* inflection (voiced or unvoiced vowel), vowel position, and vowel height. Furthermore, in an experiment on artificial language learning of bigrams, they found that participants used phonological cues when distributional cues were weak or absent. Since grammatical categories can encompass not only syntactically disparate words but also semantically disparate words, this might indicate that sound-meaning correspondences are a boon to language learners, especially in low-frequency cases.

Farmer, Christiansen, and Monaghan (2006) expanded the research on phonological diagnostics for lexical category membership begun in Monaghan et al. (2005). They performed a regression analysis on over 3,000 monosyllabic English words that significantly associated certain phonological features with an unambiguous interpretation as either a noun or a verb. An associated series of experiments demonstrated reaction time, reading time, and sentence comprehension advantages for phonologically "noun-like nouns" and "verb-like verbs."

Bergen (2004) used a morphological priming paradigm to test whether there was a processing advantage for words containing phonaesthemes over words that shared only semantic or only formal features, or which contained "pseudo-phonaesthemes." He found a difference in reaction times between the phonaestheme condition and the other three conditions by comparing primed reaction times

[1] For example, Subject-Verb-Object word order implicates syntax; persons, places, and things (nouns) are semantically different from actions and states of being (verbs).

to RTs to the same words in isolation, drawn from Washington University's English Lexicon Project. He demonstrated both a facilitation effect for word pairs containing a phonaestheme and an inhibitory effect for word pairs in which the prime contained a pseudo-phonaestheme. His use of corpus-based methods (in this case, Latent Semantic Analysis: Landauer, Foltz & Laham, 1998) was limited to ensuring that his list of words used in meaning-only priming pairs did not have any higher semantic coherence than the list of words used in phonaestheme priming pairs.

Finally, Hutchins (1998, Study 1 and 2) examined participants' intuitions about 46 phonaesthemes drawn from nearly 70 years of speculation about sound-meaning links in the literature. In her studies, participants ranked phonaestheme-bearing words' perceived coherence with a proposed gloss or definition meant to represent the meaning uniquely contributed by the phonaestheme. Participants also assigned candidate definitions to nonsense words containing phonaesthemes at rates significantly above chance, while words without phonaesthemes were assigned particular definitions at rates not significantly different from chance. She also examined patterns internal to phonaesthemes: strength of sound-meaning association, regularity of this association, and "productivity," defined as likelihood that a nonword containing that phonaestheme will be associated with the definition of a real word containing that phonaestheme.

In this paper, we chose a different approach to the study of phonaesthemes than previously demonstrated in the literature. Where previous approaches relied on intuition or a structured lexicon to gather words with candidate phonaesthemes in them, we use a corpus analysis of texts available through Project Gutenberg (Lebert, 2005). In the following experiment, we examine 47 distinct groups of words bearing candidate phonaesthemes from a large corpus using a statistical method based on Latent Semantic Analysis.

Previous studies of phonaesthemes relied on the intuitions of participants to verify the sound-meaning relationships of interest (e.g., Hutchins, 1998; Bergen, 2004). These methods are at their best when testing only a limited number of phonaesthemes. As a result, such studies have often constrained their examination to only a handful of phonaesthemes. Even the most extensive of these works, Hutchins (1998), who identified over 100 phonaesthemes previously indicated in the literature, uses only 46 of them in her experiments. In contrast with the experimental methods employed by Hutchins (1998), Bergen (2004), and others, we used a computational method based on LSA to explore sound-meaning relationships such as those exhibited by phonaesthemes.

One of Latent Semantic Analysis' most useful features is that it can be used to compute and compare semantic vectors of words and phrases. We use this feature to compare the semantic relatedness clusters comprised of words that share a phonaestheme to clusters comprised of words chosen at random from the entire corpus. Because the phonaestheme as a construct necessarily involves a partial overlap in meaning beyond that generally found in language, we hypothesize that words sharing a phonaestheme would exhibit greater semantic relatedness than words chosen at random from the entire corpus. This computational approach to the problem has two distinct advantages over the experimental methods commonly found in the literature. First, this method is objective and does not rely heavily on intuition on either the part of the experimenter or participants[2]. Second, it is possible to use the method to test a large number of candidate phonaesthemes without requiring us to probe each participant for hundreds of linguistic intuitions at a time.

The Experiment

Method

For our computational model we used Infomap (http://infomap-nlp.sourceforge.net/; Schütze, 1997), a variant of Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer et al., 1998). Infomap represents words as vectors in a multi-dimensional space whereby the distance between the words is inversely proportional to their semantic similarity. This space is constructed by reducing the number of dimensions of a matrix that records the frequency of co-occurrence between content words in the corpus through the application of a statistical method known as *singular value decomposition*. For the purposes of Infomap, two content words are said to co-occur if they are found within a specific distance from each other (i.e., for Infomap the co-occurrence frequency of *swim* and *water* could depend on how many times the word *swim* appears within 15 words of *water*). This results in a space in which the vectors for words that frequently co-occur are grouped closer together than words that rarely co-occur within the analysis window. As a result, words which relate to the same topic, and can be assumed to have a strong semantic relation, tend to be grouped together. The semantic relationship between two words can then be measured by correlating the vectors representing those two words within the semantic space.[3]

Leveraging this property of semantic spaces allows us to test the hypothesis that pairs of words sharing a phonaestheme are more likely to share some aspect of their meaning than pairs of words chosen at random. We tested whether this was true for any specific candidate phonaestheme using a Monte Carlo analysis. We first identified all of the words in the corpus that shared a

[2] At present the experimenters choose which phonetic clusters to test, meaning that intuition is still part of the process. However, whether or not any phonetic cluster qualifies as a valid phonaestheme is entirely statistically determined.

[3] This correlation is equivalent to calculating the cosine of the angle formed by the two vectors.

conjectured phonaestheme.[4] This resulted in a word cluster representing each candidate phonaestheme. Next we performed two separate Monte Carlo analyses. The first analysis averaged the semantic relationship of 1000 instances of word pairs chosen at random from the cluster. This was designed to measure the overall semantic relationship of words within the cluster. A second analysis tested the statistical significance of this relationship by running 100 t-test comparisons. Each of these tests compared the relationship of 50 pairs of words chosen at random from the conjectured cluster with 50 pairs of words chosen at random from a similarly sized cluster that was randomly generated from the list of 20,000 words for which Infomap computed vectors. We recorded the number of times these t-tests resulted in a statistically significant difference ($\alpha = .05$). Both of these analyses were performed 3 times for each conjectured phonaestheme and the median value for each run was used as the final result.

Materials

We used a corpus based on Project Gutenberg (http://www.gutenberg.org/). Specifically, we used the bulk of the English language literary works available through the project's website. This resulted in a corpus of 4034 separate documents consisting of over 290 million words. Infomap analyzed this corpus using default settings (a co-occurrence window of 15 words and using the 20,000 most frequent content words for the analysis) and its default stop list.

The bulk of the candidate phonaesthemes we used were taken from the list used by Hutchins (1998) with the addition of two possible phonaesthemes that seemed interesting to us. We also included several letter combinations that we thought were unlikely to be phonaesthemes in order to test the method's capacity for discriminating between phonaesthemes and non-phonaesthemes. Overall we examined 50 possible phonaesthemes. Of these, 46 were taken from the list Hutchins' used in her first study[5], two were candidates that we considered to be plausible phonaesthemes (*kn-* and *-ign*), and for the last two we chose phonemic sequences we thought were unlikely to be phonaesthemes (*br-* and *z-*), yielding a final list of 47 candidate phonaesthemes.

For each phonaestheme we collected all of the instances of that phonaestheme from the 20,000 most frequent content words based on an orthographic match. For each individual word stem, all but one occurrence of the stem were removed from the list (e.g., from the list for the phonaestheme '-asp' we removed the words 'clasped' and 'clasps' and retained the word 'clasp'). Preference was given to retaining the stem itself whenever it was available in the list. We also removed all words whose pronunciations were inconsistent over the same orthographic representation (e.g., the word 'touch' was removed from the list of words for the phonaestheme '-ouch'). A sample list of words is given in Figure 1.

Knack	Knapsack	Knave
Knee	Kneel	Knew
Knife	Knight	Knit
Knob	Knock	Knoll
Knot	Knuckles	

Figure 1 – List of words beginning with the phonaestheme *kn-*

Results

Our two measures, the average strength of the semantic relationship and the overall frequency of statistically significant t-test comparisons, were highly correlated ($r = 0.93$). This indicates that our method is reliable and that its results are reproducible. Because of this high correlation, our analysis is focused on the frequency of statistically significant t-tests, as this analysis is likely to apply equally to the strength of the semantic relationship and is easier to interpret from a statistical perspective.

To determine whether a conjectured phonaestheme was statistically supported by our analysis we compared the overall frequency of statistically significant t-tests with the binomial distribution for our α (.05). After applying a Bonferroni correction for performing 50 comparisons, the threshold for statistical significance of the binomial test was for 14 t-tests out of 100 to turn out as significant, with a frequency of 13 being marginally significant. We therefore judged significance frequencies (*#Sig* below) of 15 and higher to indicate a phonaestheme for which we had statistical evidence. We judged significance frequencies of 13 and 14 to indicate a phonaestheme for which we had only marginal statistical support. A list of the results for each of the tested phonaesthemes can be found in Appendix A.

Among Hutchins' original list of 46 possible phonaesthemes, we discovered 27 statistically reliable phonaesthemes and one marginally reliable phonaestheme. Overall our results were in line with the empirical data collected by Hutchins. By way of comparing the two datasets, *#Sig* and Hutchins' average rating measure were well correlated ($r = .61$). Neither of the unlikely phonaestheme candidates we examined were statistically supported by our test ($\#Sig_{br-} = 8$; $\#Sig_{z-} = 6$), whereas both of our newly hypothesized phonaesthemes were statistically supported ($\#Sig_{kn-} = 41$; $\#Sig_{-gn} = 27$).

Interestingly, there was a negative correlation ($r = -0.44$) between the number of tokens for a given phonaestheme and its significance frequency. However, it is important to note that this correlation is not unique to our method as it is also

[4] It is important to note that due to the nature of a written corpus, the match was orthographical rather than phonetic. However, in most cases the two are highly congruent.

[5] After examining our corpus we decided to drop three of Hutchins' list of phonaesthemes ('str_p', 'sp_t', and '-isp') because each of them had 3 or fewer types in our corpus and were therefore not suitable for statistical analysis. It should be noted that two of these three ('str_p' and '-isp') also had only 3 types according to Hutchins.

evident in the results reported by Hutchins (e.g., r = -0.62 between #Type and the average rating in Hutchins' study 1).

Discussion

We successfully used our computational method to verify phonaesthemes using a statistical corpus analysis. These results were congruent with the empirical data collected by Hutchins, suggesting that this statistical method can be used as a tool to examine the validity of conjectured phonaesthemes. Unlike previous work, our model can be used to directly test whether a cluster of words containing a phonaestheme is more semantically similar than would be expected by chance. While we successfully applied this test to discriminate between phonaesthemes and pseudo-phonaesthemes, at present our method does not identify what specific semantic content is carried by the identified phonaestheme. Incorporating statistical methods designed to identify the topic of a given text, such as those suggested by Griffiths and Steyvers (2002) and Blei, Ng, and Jordan (2003) may allow us to extract the specific semantic content associated with the phonaestheme, using the same corpus in which we observed the meaning vectors that identified these phonaesthemes.

It is interesting to note that the most frequently cited phonaesthemes also exhibited an exceptionally high level of support (e.g., $\#Sig_{\text{gl-}} = 96$; $\#Sig_{\text{-ump}} = 75$). This supports the common intuition about these phonetic groups' internal sound-meaning relationship and suggests that intuition tends to pick out the strongest phonaesthemes rather than weaker ones. Because of this tendency, it is likely that there is an inherent bias toward specific "easy to find" phonaesthemes in the literature. For instance, it is possible that phonaesthemes can also be infixes, but all of the phonaesthemes identified so far have been either prefixes or suffixes. In order to more rigorously test the effect of phonaesthemes on processing, a better method for identifying phonaesthemes of various degrees of strength and semantic coherence is required. Our computational method can be adapted for such use within a given corpus or across several corpora, and is therefore more suitable for the task of phonaestheme detection than any previous method of which we are aware.

At the same time, it is important to note that our method does not always validate the intuitions of previous authors. For example, we found statistical support for only 27 of the 46 phonaesthemes Hutchins examined. One possible reason for this is that some phonetic clusters drawn from our corpus had low internal semantic coherence overall, but are found in a small subset of words that are highly coherent with each other. This is especially likely in larger clusters (e.g., *gr-* which had 66 tokens). However, it is important to remember that phonaesthemes are defined as a sound or cluster of sounds that acts as a carrier for semantic content. If the number of words that share the phonaesthetic meaning is only a small subset of the number of words exhibiting the phonetic cluster then it is possible that the shared meaning is not due to a phonaestheme, but instead a common etymological root or a shared vowel quality within the subset. Indeed, it is likely that a large enough set of words, even if random, will contain a subset that shares some semantic content.

It is also possible that differences between our results and those reported elsewhere may be attributable to the age of our corpus. While the Gutenberg corpus is large, it is also drawn largely from works composed in the 19th century. It would therefore be interesting to replicate this work with more recent corpora like the British National Corpus (BNC) or Touchstone Applied Science Associates (TASA) corpus.

Finally, while the method we present in this paper is useful for the examination and identification of phonaesthemes within texts such methods are unlikely to afford sufficient insights into the processes that underlie the association between phonetic structure and semantic content. However, because our method can be used to identify new phonaesthemes and to compare the relative strength of various phonaesthemes, we hope that it will enable researchers to generate experimental designs that can examine the relationship between phonetic form and semantic content.

In concurrent research (Otis & Sagi, in prep), we are examining the effect of phonaesthemes on sentence processing and paraphrasing. The stimuli are nonsense nouns bearing phonaesthemes embedded in sentences that are either congruent or incongruent (determined by an LSA document-to-term analysis) with a real word bearing that phonaestheme. Participants are asked to read these sentences and then write a paraphrase of each. Preliminary analysis indicates that paraphrase typing latency in the presence of a congruent phonaestheme is significantly less than in the presence of a phonaestheme whose meaning is incongruent with the sentence paraphrased.

This study provides a necessary counterpart to the evidence from our computational method for detecting phonaesthemes. Behavioral tests of the effect of phonaesthemes on language processing can show us that the patterns that link sound and meaning in large corpora are not merely metalinguistic intuitions or artifacts of applying a mathematical method to large corpora, but are also psychologically real for language users.

References

Bergen, B. (2004). The Psychological Reality of Phonaesthemes. *Language*, 80(2), 291-311.

Blei, D., Ng, A. Y., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blust, R. (2003). The phonestheme ŋ in Austronesian languages. *Oceanic Linguistics* 42: 187-212.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103(32), 12203-12208.

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

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive society* (pp. 381-386). Mahwah, NJ: Lawrence Erlbaum Associates.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88-96.

Hutchins, S. S. (1998). The psychological reality, variability, and compositionality of English phonesthemes. *Dissertation Abstracts International*, 59(08), 4500B. (University Microfilms No. AAT 9901857).

Jakobson, R., and Waugh, L. (1979). *The sound shape of language*. Bloomington: Indiana University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Lebert, M. (2005). *Project Gutenberg, from 1971 to 2005.* Retrieved January 22nd, 2008 from http://www.etudes-francaises.net/dossiers/gutenberg_eng.htm

Monaghan, P., Chater, N., & Christiansen, M. (2005). The differential role of phonological and distributional cues in grammatical categories. *Cognition*, 96(1), 143-182.

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28, 225-252.

Otis, K., & Sagi, E. (in preparation). Coherence is not just semantic: How phonaesthemes facilitate language processing.

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

Schütze, J. (1997). *Ambiguity Resolution in Language Learning*. Stanford CA: CSLI Publications.

Wallis, J. (1699). *Grammar of the English Language*. Oxford, UK: Lichfield.

Infomap [Computer Software]. (2007). http://infomap-nlp.sourceforge.net/. Stanford: CA.

Appendix A – Detailed results

Table 1: Prefix Phonaesthemes from Hutchins (1998)

Cluster	Strength	#Sig	#Tokens
bl-	0.047	16	42
cl-	0.033	7	62
cr-	0.023	6	64
dr-	0.046	16	41
fl-	0.052	13	53
gl-	0.120	96	22
gr-	0.028	5	66
sc-/sk-	0.038	12	72
scr-	0.050	28	16
sl-	0.044	12	40
sm-	0.048	21	17
sn-	0.080	38	16
sp-	0.023	8	69
spl-	0.069	31	6
spr-	0.121	92	8
squ-	0.038	10	11
st-	0.028	9	139
str-	0.051	16	38
sw-	0.045	18	28
tr-	0.033	5	84
tw-	0.058	23	23
wr-	0.067	22	22

Table 2: Suffix Phonaesthemes from Hutchins (1998)

Cluster	Strength	#Sig	#Tokens
-ack	0.056	28	23
-am	0.064	33	17
-amp	0.011	5	9
-ap	0.060	47	18
-ash	0.052	17	14
-asp	0.204	100	4
-awl	0.074	53	6
-ick	0.067	44	18
-inge	0.018	9	4
-ip	0.064	39	20
-irl/-url	0.086	68	4
-ng	0.035	15	36
-nk	0.037	6	33
-oil	0.048	18	8
-olt	0.064	43	4
-oop	0.055	30	10
-ouch	0.001	7	4
-owl	0.161	96	6
-sk	0.029	7	15
-ump	0.095	75	11
-ust	0.028	5	17

Table 3: Additional phonaesthemes and non-phonaesthemes

Cluster	Strength	#Sig	#Tokens
kn-	0.072	41	14
-ign	0.059	27	14
br-	0.029	8	68
z-	0.011	6	8