Statistical Techniques for Detecting and Validating Phonesthemes

Scott Drellishak

Department of Linguistics

University of Washington

Seattle, WA

2006

1 Introduction

In the lexicons of many of the world's languages, there seem to exist subword patterns of sound and meaning that cannot easily be analyzed as morphemes. English, for example, has a number of words that start with the consonant cluster *gl-* and share a meaning related to light or vision, including *glimmer*, *glisten*, *glitter*, *gleam*, *glow*, and *glint*. Firth (1930) coined the term PHONESTHEME to describe such patterns.[1] In this paper, I adopt the following definition of phonestheme from Bergen (2004):

(1) [F]orm-meaning pairings that crucially are better attested in the lexicon of a language than would be predicted, all other things being equal. (2004: 293)

Even while proposing over a hundred phonesthemes in English alone, linguists have long struggled with their status in theories of natural language: whether or not they qualify as morphemes, how they are related to sound symbolism, and how to decide if they are real rather than mere coincidences in the lexicon. Researchers including Hutchins (1998) and Bergen (2004) have conducted psycholinguistic experiments intended to demonstrate that phonesthemes have psychological reality. Such experiments hold out the promise of proving that phonesthemes form some part of the mental grammar of language users; however, they rely on the researcher being able to select strong candidate phonesthemes for their experiments. The psycholinguist, in other words, is faced with the necessity of somehow selecting phonesthemes before experiments requiring significant time and resources can be conducted to validate those phonesthemes. Furthermore, although there is a long history of proposed phonesthemes in English, other less-studied languages may not share this accumulated resource. In this paper, I propose and evaluate

[1] In fact, he used the spelling *phonaestheme*, which is also sometimes spelled *phonæstheme* or *phonestheme*. The latter spelling is used here throughout except in quotations.

three statistical, language-independent methods for evaluating candidate phonesthemes that require only a dictionary of the target language in an electronic format and a computer running the necessary software.

2 Background

Researchers studying phonesthemes have, broadly speaking, addressed them in three ways: first, simply proposing particular phonesthemes and their meanings; second, trying in various ways to formalize the theoretical treatment of phonesthemes; and third, attempting to determine if, or to what extent, they are real.

Although Firth (1930: 184) coined the term phonestheme, he was not the first to notice these patterns in the English lexicon. Wallis (1653), in the section of his *Grammatica Linguae Anglicanae* devoted to etymology, which term he used in the sense of word formation as well as word origins, describes a number of phonesthemes (though not so called), including:

> Str. Sic voces a *Str* inchoatæ fortiores rei ſignificatæ vires innuunt; ut *ſtrong* fortis, *ſtrength* vires, *ſtrîve* validè contendo, *ſtrîke* percutio, *ſtruggle* luctor, *ſtretch* extendo, *ſtrain* violenter extendo, *ſtraight* rectum (quod nempe in longitudinem extenditur,) *ſtrout* tumeſco (diſtendor) quantum poſſum.
>
> Thr. *Thr* violentiorem motum innuunt: Ut *throw* projicio, *thrust* violenter trudo, *throng* conſtipo (de caterva dici folet,) *throb* violenter palpito (de corde acerrimis doloribus agitato dicitur,) *through* penitus, per totum, &c.[2] (1699: 120-121)

[2] *Str*. Thus expressions beginning with *Str* point to the strength of the powerful thing signified; for example *strong*, *strength*, *strive* 'compete strongly', *strike*, *struggle*, *stretch*, *strain* 'stretch violently', *straight* 'straight' (that which is truly extended in length), *strout* 'swell (be stretched) as far as possible'. *Thr*. *Thr* points to violent motion: for example *throw* 'throw out', *thrust* 'push violently', *throng* 'crowd together', *throb* 'beat violently' (said of a bitter heart driven by sorrows), *through* 'within, all the way through, etc.' (translation mine)

Firth (1930: 184) characterized phonesthemes as “initial and final phone groups not ordinarily recognized as having any function,” (1930: 184) He notes a group of English words beginning with *sl-* that he claims share a pejorative meaning, including: *slack, slouch, slush, sludge, slime, slosh, slash, sloppy, slug, sluggard, slattern, slut, slang, sly, slither, slow, sloth, sleepy, sleet, slink, slip, slipshod, slope, slit, slay, sleek, slant, slovenly, slab, slap, slough, slum, slump, slobber, slaver, slur, slog,* and *slate*. He writes, “The more consistently similar sounds function in situations having a similar affective aspect, the clearer their function. In this way, then, *sl* can be said to be a pejorative phonetic habit.” (1930: 185) In his view, such habits reinforce, and are reinforced by, the related meanings of the words containing them.

Firth’s treatment of phonesthemes, although seminal, is rather superficial, with only the vague and subjective (“not ordinarily recognized”) definition quoted above. Moreover, it is not clear what theoretical status Firth assigns phonesthemes. It may seem that, by calling them “phonetic habits”, he is treating them as extra-linguistic and distinguishing them from other more familiar language phenomena. This is not the case, however; Firth considers the phoneme, a linguistic phenomenon if ever there was one, to be another kind of phonetic habit. His account of phonesthemes relies on the strength of his examples to make clear what they are, leaving it to later researchers to define them in more detail.

Bloomfield (1933) discusses phonesthemes (without using the term) in a chapter on morphology. He writes, “we find clearly-marked phonetic-semantic resemblances between elements which we view as different roots,” then gives as an example the onsets in the English pronoun system:

[ð-]: *the, this, that, then, there, thith-er, thus*.

[hw-]: *what, when, where, whith-er, which, why*; modified to [h] in *who, how*.

[s-]: *so, such*.

[n-]: *no, not, none, nor, nev-er, neith-er*. (1933: 244)

It is interesting that this pattern occurs in function words; phonesthemes are typically proposed for open classes (nouns, verbs, and adjectives). Bloomfield next turns to this more familiar variety, writing "we can distinguish, with varying degrees of clearness, and with doubtful cases on the border-line, a system of initial and final *root-forming morphemes*, of vague signification," and proposing more than a dozen of them, including *fl-* 'moving light' (*flash*, *flare*), *fl-* 'movement in air' (*fly*, *flit*), and *gl-* 'unmoving light' (*glow*, *glare*). Bloomfield's analysis is more explicit than Firth's—he states clearly that, since they represent phonetic-semantic relationships, phonesthemes should be treated straightforwardly as morphemes. He admits, however, that it can be difficult to pin down their exact meaning, or even to determine if a proposed phonestheme represents a true "linguistic form", because that requires somehow evaluating, for the words in the set, their semantic similarity, "[for] which [since it] belongs to the practical world, we have no standard of measurement." (1933: 246) My aim in this paper is to provide an empirical, statistical standard for this measurement.

Although the morphemic analysis of phonesthemes has not been universally adopted, Rhodes and Lawler (1981) also maintain that phonesthemes are merely sub-syllabic morphemes, no different in principle from other morphemes. In a section analyzing English monosyllables like *stump*, *clump*, *sting*, and *cling* as made up of onset and rhyme morphemes with compositional semantics, they write, "the units which we analyze out of the monosyllable are simple morphemes…we claim that both the (internal) syntax of the monosyllabic construction and the semantic nature of the component morphemes is more limited and systematic than was previously thought." (1981: 326)

Other researchers have treated phonesthemes as a variety of sound symbolism. Jespersen (1922), after a discussion (1922: 398-9) of words that directly imitate sounds and refer either to the sound itself (e.g. *clink, cock-a-doodle-doo*) or to the originator of that sound (e.g. *cuckoo*), compares them to what he calls "words expressive of such movements as are not to the same extent characterized by loud sounds". He suggests that this latter group includes a large number of words beginning with consonant clusters ending in *-l-*, including among others *flow*, *flutter*, *fling*, *slide*, *slip*, and *glide*. (1922: 399-400) In spite of Jespersen's analysis of this as sound symbolism, the connection between the sound of these words and the meaning 'movement' seems obscure; Bolinger (1965), in support of Jespersen's analysis, asserts that such patterns must originally have had a sound-symbolic value that has been lost:

> What may have been the original sound significance of *gl* and related sounds for the eye and visual appearances would be difficult to single-out—that there was sound symbolism seems to be indicated by the great number of words that show this uniformity; yet the disappearance of the sound symbolism has not affected the vigor of the constellation… (1965: 195)

In all of these discussions and analyses of phonesthemes, the researchers have been largely silent about an important question: how can we know that the phonesthemes they propose are in some sense real linguistic phenomena, and not just coincidences in the lexicon? The list of proposed phonesthemes has grown over time by accretion, with each researcher reporting the proposals then extant in the literature, then suggesting more possibilities based on little more

than intuition. Hutchins (1998) describes her iteration of this process[3], writing, "Many of these phonesthemes had been identified by previous researchers…others were candidates for phonestheme status that did not appear previously in the literature but seemed likely to the investigator." If this methodology is applied without a standard of proof for validating phonesthemes, linguists run the risk of accepting the reality of *any* phonestheme proposed by a researcher. Consider the *cr-* phonestheme, which Bloomfield (1933: 245) suggests has the meaning 'noisy impact' (e.g. *crash*, *crack*, *crunch*). There are other English words beginning with *cr-* that have unrelated meanings (e.g. *cream*, *crawl*, *crime*, *create*, and *cruel*). Does the proportion of *cr-* words with the phonesthetic meaning support the existence of the phonestheme? Answering this question becomes increasingly challenging as the number of words with the proposed phonetic content becomes large, as for Bloomfield's proposed *j-* phonestheme, meaning 'up-and-down movement', for which he gives seven examples. Do only seven words with that meaning out of all the English words beginning with *j-* represent a pattern that is more than coincidence?

3 Validating Phonesthemes

What is needed, then, is a way to convincingly prove the existence of phonesthemes, and, furthermore, validate particular proposed phonesthemes. Two possible approaches seem promising: statistical and experimental.

3.1 Statistical Validation

Statistical approaches have the advantage of being relatively inexpensive in terms of resources and time. A simple approach such as finding all the words with some phonetic content

[3] Unlike many previous researchers, however, Hutchins goes on to test her list of proposed phonesthemes by conducting psycholinguistic experiments, which are described in more detail in §3.2.1.

and counting up the number that have the proposed phonesthetic meaning, requires nothing more than a dictionary for the language in question. Even such simple methods have only occasionally been employed by researchers, who seem content to focus on a few of the most intuitively strong examples (such as *gl-* and *fl-*), and when statistical methods have been proposed, they lack criteria for distinguishing real correlations from chance patterns in the lexicon.

Abelin (1999) discusses Swedish sound symbolism, including phonesthemes, in great detail. At one point in this discussion (1999: 87) he calculates, for 36 initial-cluster phonesthemes, the percentage of root morphemes beginning with the cluster that have the proposed phonesthetic meaning. The values range from as low as 8% to as high as 100%. In statistical terms, it is hard to argue with 100%—apparently, every root in Swedish that begins with /fn/ is pejorative—but the lower the percentage, the more doubtful the phonestheme becomes. Is 8% a surprisingly large percentage, or could it be due only to chance?

Bergen (2004), who like Hutchins performs experiments to validate phonesthemes, actually defines phonesthemes twice. His first definition is, "frequently recurring sound-meaning pairings that are not clearly contrastive morphemes." (2004: 290) This definition relies on a negative criterion, and a subjective one at that: the clarity of a particular sound-meaning pairing's status. His second, narrower definition was adopted here as (1), repeated here for convenience:

(2) [F]orm-meaning pairings that crucially are better attested in the lexicon of a language than would be predicted, all other things being equal.

This definition makes clearer how we can distinguish phonesthemes from, for example, morphemes. Since morphemes are well understood, we would predict form-meaning pairings associated with them; phonesthemes are pairings that would *not* be predicted, therefore they must

then be a separate phenomenon. It is also explicitly a statistical definition because it makes an appeal ("better attested") to frequency. To demonstrate the consequences of this definition, Bergen examines the distribution of four onsets (*gl-*, *sn-*, *sm-*, and *fl-*) in word types and tokens in the Brown Corpus, noting for instance that 38.7% of types (distinct English words) and 59.8% of tokens (occurrences of words in the corpus) that begin with *gl-* have meanings associated with light or vision. However, he examines only these four, intuitively rather strong, phonesthemes, and does not explain how high the percentages must be before we should accept their reality, referring only to the "overwhelming statistical pairings of forms like *gl-* and *sn-* with their associated meanings." (2004: 293)

3.2 Experimental Evidence

Statistical tests for validating phonesthemes may be inexpensive and straightforward to compute, but in order to finally convince ourselves that phonesthemes really form a part of the mental grammar of language users, we must make recourse to psycholinguistic experiments that demonstrate measurable effects on the comprehension or production of phonesthetic words. Hutchins (1998) and Bergen (2004) both conducted such experiments.

3.2.1 Hutchins (1998)

Hutchins (1998) describes three experimental studies. The first study measured the "variability among English phonesthemes in the regularity of their sound-meaning associations." (1998: 14-15) Fifty monolingual English speakers were asked, for 46 different phonesthemes, to rate on a seven-point scale how well each of a list of words matched the proposed semantic content of the phonestheme. The results did show variability in the strength of the sound-meaning association for the phonesthemes studied; however, the strength of the association was inversely correlated with the frequency of the phonestheme in the lexicon. The results

additionally confirmed the (perhaps unsurprising) fact that not all words with the phonetic content of a phonestheme have the associated meaning, a fact which Hutchins takes to mean that the sound-meaning associations are probabilistic. (1998: 28)

The second study tested the psychological reality of phonesthemes. In it, each participant performed one of two tasks: either they heard a nonsense word pronounced and were asked to pick one of four definitions, or they read a definition and selected one of four nonsense words. The results support the hypothesized psychological reality of phonesthemes: in both tasks, participants chose a phonesthetic match approximately twice as often as would be expected by chance. (1998: 38)

The third study tested the possibility that phonesthemes might be made up of even smaller, compositional elements. Its design was similar to the second study, except that instead of being presented with nonsense words containing a proposed phonestheme, participants were presented with nonsense words containing a *different* phonestheme that shared at least one phoneme with the proposed one. Hutchins hypothesized that, if some phonesthemes are made up of smaller compositional elements, there should be a greater-than-chance association between semantic glosses and nonsense words containing phonetically-related phonesthemes. The results for the third study seem to show some evidence of compositionality, but Hutchins points out alternative explanations for these results and writes that "[f]inal conclusions regarding the compositionality of English phonesthemes…await more systematic tests." (1998: 46)

The results of Hutchins' three studies support the reality of phonesthemes (although, as we will see below, Bergen (2004) points out some potential methodological weaknesses). Hutchins' experiments are also valuable because of the large number of phonesthemes evaluated. Moreover, in an appendix to her dissertation, Hutchins collects an extensive list of English

phonesthemes that have been proposed by previous researchers. The list includes 145 phonesthemes, both onsets and rhymes, a number of which have multiple, sometimes partially overlapping, proposed meanings. For example, she cites 12 proposed meanings for the onset *fl-* including "expressive of movement" (Jespersen 1922), "cognate of syllabic 'fall'" (Wescott 1987), and "moving light" (Bloomfield 1953).

3.2.2 Bergen (2004)

Bergen (2004) describes another experiment designed to demonstrate the psychological reality of phonesthemes. He points out that experiments (including Hutchins') that allow the participants time for reflection are flawed:

> [O]ne could still hold the position that phonaesthemes are only static, distributional facts about the lexicon, which speakers of a language can access consciously. This is problematic since essentially all normal morphological processing happens unconsciously. We know that language users are able to access all sorts of facts about their language upon reflection. People can come up with a word of their language that is spelled with all five vowel letters and 'y' in order, or a word that has three sets of double letters in a row. These abilities by themselves, though, do not lead to the conclusion that orthographic order of vowel letters in a word is a fundamental principle of implicit cognitive organization. For the same reason, subjects' ability to consciously access distributions of sound-meaning pairings in their language does not imply that those pairings are meaningful for the subjects' linguistic system. (2004: 295)

In order to avoid this problem, Bergen's experiment was designed to test his participants' unconscious language processing. The experiment was a morphological priming study in the sense of Kempley and Morton (1982), in which participants were presented briefly (150 ms) with

a prime word, then 300 ms later, asked to decide if a second, target word was a word of English or not. There were five categories of stimuli:

1. Both the prime and the target had the phonetic content (an onset) and meaning of a proposed phonestheme
2. The prime and the target shared an onset
3. The prime and the target shared some meaning
4. The prime and the target shared an onset and some meaning, but the frequency of this sound-meaning pairing was so low it could not be a phonestheme (Bergen calls these "pseudo-phonaesthemes", and mentions *crony* and *crook* as an example).
5. The prime and target were unrelated (2004: 297)

The results of Bergen's experiment show that participants processed the phonestheme pairs significantly differently from the others. They responded 59 ms faster on average when the prime and the target shared a phonestheme (category 1): 606.7 ms versus 665.3 ms for unrelated primes and targets (category 5). Pairs sharing only a meaning were also processed somewhat faster (23 ms). In the case where the prime and target shared only an onset, however, the participants' responses were actually slightly slower than the baseline (668.2 ms versus 665.3 ms). (2004: 299) These results convincingly demonstrate that, even when the experiment rules out the possibility that participants are consciously searching for relationships between words, processing speed is affected by the phonesthetic content of those words.

4 Goals

Psycholinguistic experiments can convincingly prove the psychological reality of phonesthemes, irrespective of whether we analyze them as morphemes, sound symbolism, or some other linguistic phenomenon. Unfortunately, such experiments are time-consuming, and

the number of proposed English phonesthemes collected by Hutchins (1998) is large. It is desirable that there should be a simple, inexpensive procedure for validating proposed phonesthemes. Adopting the statistical definition of phonesthemes of Bergen (2004) allows us to characterize them regardless of how they are analyzed, and suggests the possibility of statistical criteria for selecting candidate phonesthemes:

(3) a. The phonesthetic meaning must be associated with the proposed phonetic content of the phonestheme with greater than chance frequency.

b. The pattern being proposed as a phonestheme must not be explainable by any other linguistic phenomenon; in particular, it must not be due to a known etymon or morpheme.

It is important to note that a method based on such statistical criteria will be prone to false positives. Correlations within the lexicon of a language between sound and meaning might be due to the presence of other well-understood linguistic phenomena, particularly morphemes and etyma. Any method for detecting phonesthemes must address the possibility that a detected sound-meaning correlation is a morpheme, more or less distorted by phonological or morphophonological processes. We would expect, for example, that *un-* is correlated with a meaning related to negation. Etyma present a similar problem. For example, we would expect headwords containing the Latin root *-viv-* to be highly correlated with a meaning of ‘life’. Both of these kinds of false positives must be ruled out somehow, perhaps by human supervision.

It is also important to note that no statistical method can truly prove the existence of a phonestheme. There is every reason to believe that human languages are imperfect systems—even if we can show statistically that it *would* be more efficient if the mental lexicons of speakers of some language were organized to take account of a proposed phonestheme, that is no

guarantee that they *are* so organized. Ultimately, only psycholinguistic experiments like those of Bergen and Hutchins can show that phonesthemes really are part of speakers' grammars.

5 Methodology

In order to evaluate whether a phonesthemes is associated with a meaning with greater than chance frequency, we must decide across which domain the frequencies are to be measured. There are two obvious candidates: frequency within the lexicon and frequency in some corpus. In the techniques described in the following sections, I have focused on frequency in the lexicon because that is the domain to which phonesthemes have been assumed to belong in the literature. Previous researchers have compared them to morphemes (Bloomfield 1933, Rhodes and Lawler 1981) and to phonemes (Firth 1930), for example, both of which exist in contrasting paradigms in the mental grammars of speakers and not in a particular assemblage of words in a corpus. It is possible that the other approach—that is, to consider the frequency of phonesthemes within some corpus—may have some utility, but that is outside the scope of this paper.

Implementing a method for detecting phonesthemes computationally requires a dataset for the language being studied. Ideally, this would consist of a database containing complete details of the phonetic and semantic content of the lexical items being studied. The methods described here use an English dictionary, the freely available 1913 edition of Webster's dictionary, as a substitute for such an ideal database. The orthography of headwords is used as a proxy for pronunciation—though admittedly the mapping between the two is less than straightforward in English—and the presence or absence of words in definitions is used as a proxy for meaning. These assumptions allow the use of existing resources rather than the costly and time-consuming creation of novel ones.

5.1 Latent Semantic Analysis (LSA)

All the methods described here are varieties of Latent Semantic Analysis (Deerwester et. al. 1990). In LSA, a set of documents is described by a term-document matrix. Each row in this matrix is a vector of counts of words occurring in one of the documents, also known as a word feature vector; each column therefore contains the counts, in all documents in the set, for a particular word. For the purposes of phonestheme detection, the definition of each headword[4] in the dictionary is treated as a separate document. The first detection method described here is based on DOCUMENT CLUSTERING, in which documents (or rather, their corresponding rows in the term-document matrix) are grouped into clusters based on similarities in their word feature vectors. The other two detection methods described here fall into the category of DOCUMENT CLASSIFICATION, which involves the discrimination, based on their word feature vectors, between two or more sets of documents.[5]

5.2 Clustering

One LSA technique that might be used to detect phonesthemes is clustering, in which similar rows in the term-document matrix, which represent similar documents, are grouped algorithmically into clusters. The clustering method for phonestheme detection is as follows. First, take the word feature vectors from two or more sets of definitions and put them into a single large matrix, then apply automatic clustering to group definitions that have similar distributions of words. If one or more of the classes contains a phonestheme then, given the right settings for the clustering algorithm, there should be a cluster that contains a higher fraction of its

---

[4] In the following discussion, the term *headword* will consistently be used to refer to a word with a definition, while the words within the definition will be called *definition words* or simply *words*.

[5] Bergen (2004: 301) mentions another LSA technique he calls the pairwise comparison function, which measures similarity between the contexts in which two words appear. He uses it to address concerns that his phonestheme prime-target pairs might have been more closely semantically related than the other categories (which they turn out not to be), rather than using it to validate his candidate phonesthemes.

definitions. Clustering should work, in principle, because words associated with a phonestheme's meaning should occur with greater than chance frequency in the definitions of headwords containing that phonestheme. The advantage of the clustering approach is, if it can be made to work, more than one proposed phonestheme can be tested in a single pass.

Here is how the clustering method would work in an ideal case. Suppose we applied automatic clustering to three sets of definitions A, B, and C. All of the definitions in A share some orthographic feature (e.g. they all begin with *gl-*) and 30% of them have a phonesthetic meaning. B is similar to A, except that it contains a different candidate phonestheme. C is a set of randomly selected definitions. A hypothetical ideal result would look like this:

	Cluster 1	**Cluster 2**	**Cluster 3**
A	30%	0%	70%
B	0%	30%	70%
C	0%	0%	100%

Cluster 1 contains all the phonestheme words from A, Cluster 2 contains all the phonestheme words from B, and Cluster 3 contains all the non-phonestheme words, including all of C. Of course, the results in practice are unlikely to be so categorical. Other competing sound-meaning associations, including etyma and morphemes, will tend to cause non-phonestheme clusters to occur. Therefore, the clustering method's results will need to be evaluated by a human, who by examining the characteristic words for each cluster—that is, the words most strongly associated with the cluster, as reported by the clustering software—can determine if that cluster is associated with a proposed phonestheme's meaning. If settings for the clustering algorithm could be found that consistently produce correctly clustered results for known phonesthemes (such as *gl-* and *sn-*, which were validated by Bergen (2004)), then in principle it

should be possible to apply this technique to automatically-generated candidate phonesthemes in order to find phonesthemes without any human intervention.

5.3 Document Classification

Another LSA technique that might be used to detect phonesthemes is document classification, in which a statistical model is used to decide which of several classes a document belongs to. Document classification techniques can be applied to phonestheme detection in the following way. First, select from the dictionary all the definitions of headwords that match the orthographic (phonetic) content of the proposed phonestheme. Next, select a random set of definitions from the dictionary. Now consider the distribution of words that occur in the various definitions, looking for words that are highly correlated with one set or the other—or, to put it another way, words that would be very informative when trying to classify definitions as belonging to one set or the other. If the most highly correlated (or most informative) words have meanings similar to the proposed phonesthetic meaning, it would suggest the phonesthetic sound-meaning pattern is real. It is important to note that while the methods described here are based on and inspired by the mathematical methods used to perform document classification, classifications of documents are never actually performed. Moreover, because the classification methods rely on calculating a "score" for each definition word rather than on dividing definition into clusters, all definitions in each definition set will be treated as a single large document for convenience.

5.3.1 Relative Word Frequency (RWF)

A straightforward method of estimating which definition words are correlated with a particular phonestheme makes use of the frequencies of the definition words. Suppose we have a set of definitions that might contain a phonestheme. The frequency of a word in the definition

set is defined as the number of times it occurs divided by the total number of word tokens in the set. We can also calculate the word frequencies for the dictionary as a whole—that is, the set of all definitions. Now we have, for each definition word, two frequencies, one for the proposed phonestheme and one for the whole dictionary. The ratio of these two values (frequency in the phonestheme set divided by frequency in the whole dictionary) is the RELATIVE WORD FREQUENCY, and it tells us which words occur more frequently on average in the phonestheme set. If a phonestheme is real, we would expect that words with the highest RWF to be words associated with the phonesthetic meaning.

5.3.2 Mutual Information (MI)

Another way to determine which definition words are associated with a phonestheme is to calculate their MUTUAL INFORMATION, a measure of how much one random variable predicts another. Mutual information is defined in terms of the ENTROPY of the variables. According to the information-theoretic definition of Shannon (1948), entropy is the amount of information produced by a random process. For a probability distribution p, the entropy H is defined by the following formula:

(4) $$H = -K\sum_{i=1}^{n} p_i \log p_i$$ (Shannon 1948)

(Where the constant K has only to do with the choice of units.) Mutual information, in turn, is defined in terms of entropy. Intuitively, mutual information is a measure of how much information knowing the value of one random variable tells us about the value of another. For two random variables X and Y the mutual information $I(X;Y)$ is defined by the following formula:

(5) $$I(X;Y) = H(X) + H(Y) - H(XY)$$ (Fano 1961: 48)

Note that mutual information is symmetrical—that is, $I(X;Y) = I(Y;X)$. The units of mutual information (and of entropy) are determined by the base of the logarithm; when the logarithm is base two, for example, the each unit of MI is equal to one binary digit, or one bit.

Recall that we are applying the mathematical tools of text classification to the problem of phonestheme detection. To this end, we can define the mutual information between the class of a document (represented by the variable C) and the presence or absence of a particular target word in the document (represented by the variable W_t) using the following formula:

$$I(C;W_t) = H(C) - H(C \mid W_t)$$
$$= \sum_{c \in C} \sum_{f_t \in \{0,1\}} P(c, f_t) \log \left(\frac{P(c, f_t)}{P(c)P(f_t)} \right) \quad \text{(6)}$$

(McCallum and Nigam 1998: 3)

All of the values in (6) can be estimated empirically. In this method, there will always be two classes, one of which corresponds to the definitions of a proposed phonestheme, and the other to all the definitions in the dictionary. $P(c)$ is number definition words in definitions of class c divided by the total number of definition words; $P(f_t)$ is the number of occurrences of the target word divided by the total number of definition words; and $P(c, f_t)$ is the number of occurrences of the target word in definitions of class c divided by the total number of definition words. The resulting mutual information value tells us how informative the appearance of a particular word in a definition is toward classifying the definition as part of one class or the other—to put it another way, the MI of a definition word tells us how characteristic that word is of one set of definitions or the other, with high-MI words being more strongly associated with a single set and low-MI words associated with both sets.

To use MI to validate a phonestheme, then, we use the following procedure. First, we create two classes of definitions: one containing candidate phonestheme words, and the other

containing all definitions in the dictionary. Next, we calculate the MI between each definition word and the classification, then sort the words and examine the ones with the highest mutual information. If the phonestheme is real, then some or all of the words near the top of the sorted MI list should have meanings associated with the proposed phonesthetic meaning.

5.4 Data and Tools

The dictionary used as a lexical database was the 1913 edition of Webster's Dictionary, which is freely available online (Porter et. al. 1913). It contains about 110,000 headwords, of which about 53,000 have etymologies. It is in an SGML format that I reduced to plain ASCII, with all markup, punctuation, and capitalization removed. Some definitions with odd or complex formatting were discarded in this process, so the final ASCII dictionary contained 92,466 definitions and 48,468 etymologies. Some decisions had to be made during this conversion that might have had an effect on the results; in particular, all senses of a each headword (e.g. *bat* meaning 'a wooden club' and *bat* meaning 'a part of a brick') were collapsed into a single definition, but different headwords with the same spelling (e.g. *bat* meaning 'a wooden club' and *bat* meaning 'a small flying mammal') were not collapsed.

All the methods described here used the `rainbow` program, which provides a command-line interface to the BOW toolkit (McCallum 1996). It was used to train Naïve Bayes classifiers on various sets of definitions. The classifier was actually never used, but the statistics collected by `rainbow`, including the term-document matrix, were necessary for the clustering method, which was performed using the `vcluster` program, a part of the CLUTO toolkit (Karypis 2003). The document classification methods (MI and RWF) involved further processing of the statistics contained in the term-document matrix; in particular, the MI method relied on a feature

of `rainbow` that prints out the mutual information between the top *n* words and the classification.

5.5 Feature Selection

An important step in the development of statistical models is feature selection, in which the developer decides which variables should be modeled. In the techniques being described in this paper, beyond the initial decision to treat each definition as document to be classified, further feature selection was performed—or, more precisely, feature *exclusion* by filtering out definition words that tended to produced false positives in preliminary tests.

As mentioned above, morphemes and etyma are potential problems for the approach described in this paper. Morphemes such as the prefix *un-* have a similar distribution and appearance to many candidate phonesthemes and are associated with a particular meaning, but they are not phonesthemes. Etyma like the Latin root *-viv-* 'life' ought to be similarly correlated with words found in definitions. It is desirable to reduce the chance of a morpheme or etymon being detected as a phonestheme, so some feature selection (i.e. filtering) was done to reduce the chance of such false positives.

The filters were developed by repeatedly applying the mutual information method to two phonestheme sets: the **sn** set, containing the definitions all headwords beginning with orthographic *sn-*, and the **gl** set, containing all headwords beginning with *gl-*. After each application, the results were examined for classes of words having high mutual information but not associated with the phonesthetic meaning. Filters were written to remove such words, the filters were applied, and the process repeated. The result was three filters: the ETYMON FILTER, the PATTERN FILTER, and the STOPWORD FILTER.

The etymon filter removed, from the definition of each headword, any definition word that also appeared in the etymology. This was intended to prevent false positives due to etyma. It is potentially very powerful—if the source dictionary's definitions and etymologies were both written using a restricted vocabulary, and an etymology was included for every word whose etymology was known, this filter could suppress most or all etymology-related definition words that might appear to be phonesthetic meanings. Unfortunately, the freely available dictionary used was not so perfectly consistent. For example, the 1913 Webster's definition of *lutose* is 'covered with clay; miry', but its etymology is [L. *lutosus,* fr. *lutum* mud], so this filter would be unable to rule out the word *clay* as being related to an etymon. Similarly, while base forms such as the headword *chaos* have an etymologies, derived forms such as *chaotic* do not, blunting the effectiveness of this filter.

The pattern filter removes from each definition any words that match the orthographic content of the phonestheme being evaluated. So, for example, if we are evaluating *gl-*, all definition words beginning with *gl-* are removed. This is intended to prevent words like *snow* and *glass*, both of which appear quite often in their respective phonestheme sets, from being detected as phonesthetic meanings simply because they occur often in examples within their definition sets. This filter also serves to remove component morphemes of compound and derived headwords (e.g. *snowball*, *glassy*). This pattern, where a whole word in a definition occurs in the headword, is extremely unlikely to be an example of a phonestheme—if, for example, we find *snow* occurring often in the definitions of headwords like *snowball* and *snowy*, we have discovered a root morpheme, not a phonestheme. It should be noted that the use of this filter is not without cost—for example, a plausible meaning of the phonestheme *bl-* is 'blow', but *blow* would be removed from all definitions by the pattern filter.

The stopword filter removes a set of very commonly occurring words from all definitions. In the 1913 Webster's dictionary, the definitions associated with several parts of speech very often contain characteristic turns of phrase: "of or pertaining to" is often used with adjectives, "manner" often appears with adverbs, and so on. Although these words occur very frequently, they do not have any relation with phonesthetic meanings. These stopwords were especially problematic for the clustering method because their presence tended to overwhelm any phonesthetic relationships between words, instead causing it to produce clusters containing the various parts of speech. The stopword filter therefore removes the following definition words:

(7) *word*, *quality*, *pertaining*, *consisting*, *relating*, *state*, *manner*, *common*, *called*, *resembling*, *act*, *action*, *kind*, *genus*, *genera*, *species*, *quantity*

6 Results

I report below the results of all three techniques (clustering, mutual information (MI), and relative word frequency (RWF)), using all three of the filters described above.

6.1 Clustering Results

The clustering method was unsuccessful at detecting or validating phonesthemes. In general, the clustering results were unaffected by different choices of options to CLUTO's `vcluster` program, with the exception of two. First, agglomerative clustering, regardless of the other option settings, always produced one very large cluster with only a handful of definitions in the other clusters; therefore, divisive clustering was used exclusively in generating these results. Second, varying the number of clusters, from a value equal to the number of definition sets being evaluated up to 100 or so, produced significantly different results that are explored in more detail below.

With all the feature selection filters in place, the following results were obtained using the clustering method to compare both the definitions of headwords beginning with *sn-* and with *gl-* to a random set of definitions:

(8) **sn vs. random:**

	Cluster 1	**Cluster 2**	**Unclustered**
sn	63 (37%)	73 (42%)	34 (20%)
random	1379 (34%)	2038 (50%)	616 (15%)

(9) **gl vs. random:**

	Cluster 1	**Cluster 2**	**Unclustered**
gl	139 (38%)	161 (44%)	65 (17%)
random	1285 (31%)	2136 (52%)	612 (15%)

These results do not show the sort of categorical difference between the definition sets that would imply positive results. Furthermore, examining each cluster's characteristic definition words showed none that were at all related to the proposed phonesthetic meanings.

As mentioned above, it is possible to increase the number of clusters above two, in the hope that, if some stronger inter-headword relationship (e.g. part of speech) is overwhelming the desired phonesthetic relationships, a greater number of clusters might allow weaker phonesthetic relationships to form a cluster. Values of 5, 10, 20, and 50 clusters were tried with the *sn-* definitions. Finally, in the 50-cluster run, there appeared a cluster whose descriptive words were *sound*, *nose*, *noise*, *utter*, and *air*, and which contained the definitions of the words *snap*, *sneer*, *sneeze*, *sniff*, *sniffing*, *sniffle*, *snite*, *snivel*, *snively*, *snoring*, *snort*, *snot*, *snuff*, and *snuffle*.

Unfortunately, this method is fatally flawed. Increasing the number of clusters allows words with finer and finer lexical relationships to be divided into separate clusters—as more clusters become available, groups of words that were previously grouped together can split into

two smaller clusters. In fact, words with *any* relationship would eventually be grouped into their own cluster (so long as they were not distributed into multiple clusters at some previous phase of the divisive algorithm, since clusters never merge). In the 50-cluster case above, then, we have steadily increased the number of clusters until all or most of the headwords with *nose* in their definitions fallen into a single cluster. What has been proven? Only that there is some relationship between the *nose* definitions, but we knew that already: they all contain the word *nose*. Crucially, this does not show that the *sn-* form and the *nose* meaning co-occur *with greater than chance frequency*.

In order for the clustering approach to work, we would need a way either to discount other sorts of lexical relationships (perhaps using some very smart filters) or to magnify the lexical relationships associated with the phonesthemes—this would let us use only two clusters (or perhaps a slightly larger, but still strictly bounded, number of clusters) to test proposed phonesthemes. Unfortunately, no such methods are known.

6.2 Relative Word Frequency Results

Ranking definition words by relative word frequency was also unsuccessful. When the definitions for the candidate phonestheme *sn-*, for example, are compared with the entire dictionary (with all filters applied to both sets), the 40 definition words with the highest RWF are:

(10) *raley, avulsion, antirrhinum, neishout, whiningly, leucoium, alice, unstained, nemichthys, plectrophenax, colubrina, plumieria, lutjanus, sanil, nop, albocoronata, crossly, ptarmica, serpentium, swaging, galanthus, testily, wireloop, neb, inssinuate, horsed, hyemalis, vernum, ravallia, microchra, adderstongue, knobstick, trumpetwood, bentup, ruellia, impulsively, scrrophulariaceous, ophioxylon, avalanche,* and *olympus*

Furthermore, the RWF value for all of these words is exactly the same—about 514.04. The RWF values are equal because each word occurs exactly once in all the definitions in the dictionary. Its RWF is therefore equal to the total number of definition word tokens in the dictionary divided by the number of definitions word tokens in the *sn-* set, or 1,565,762 divided by 3046.

These results make the RWF method unsuitable for validating phonesthemes for two reasons. First, notice that none of the words in the set is related to the meaning of the phonestheme *sn-*, namely 'nose', whose psychological reality has been validated by both Hutchins (1998) and Bergen (2004). The definition word *nose* unfortunately had an RWF score of only about 102, placing it 145th on the list. This is still rather high given that there are 69,237 distinct definition words in the sets after filtering, but this method would not be very convenient or convincing if a researcher had to ignore more than 99 our of every 100 words it produced. Second, the fact that a large number of words that occur exactly once all have equal RWF values greatly diminishes this method's discriminative power. If the items at the top of the RWF list are simply the words that occur once, and they have no relationship to the phonesthetic meaning, the RWF method is unworkable.

6.3 Mutual Information Results

In contrast to the RWF method, the mutual information method showed promising results in testing. It was therefore applied to all 46 of the phonesthemes[6] tested by Hutchins (1998), a set that also includes the two phonesthemes tested by Bergen (2004). For most of these phonesthemes, definition words associated with the phonesthetic meaning appeared near the top of the list sorted by MI score.

---

[6] Some of these candidates are suspiciously orthographic rather than phonetic. For instance, *wr-* and *-owl* both exclude some headwords that are pronounced the same (e.g. *wring/ring*, *fowl/foul*).

To see how the method worked, consider these four phonesthemes evaluated by Hutchins:

(11) *sn-* "related to the nose, or breathing; or by metaphorical extension to snobbishness, inquisitiveness (sneeze, snout, snoop)"

st- "something firm, upright, regular, or powerful; or forceful linear motion (stab, stand, stiff)"

spr- "to radiate out from a point or to be elongated (spray, sprawl, spread)"

-Vng "a sharp, quick, or oscillating movement producing a ringing sound or sensation; or the sound produced by such an action (bang, clang, ring)"

(Hutchins 1998: 66-69)

Below are listed the top 20 definition words, sorted by MI, for the above four phonesthemes. Words that are associated with the phonesthetic meaning are in boldface:

(12) *sn-*: ***nose***, *sharp*, *reprimand*, *seize*, ***contempt***, *short*, *bite*, *with*, *laugh*, ***nasal***, *angry*, *check*, *air*, *nip*, *catch*, *fellow*, *mucus*, *surly*, *rebuke*, *mean*

st-: *to*, ***firm***, ***fixed***, *in*, ***upright***, *vessel*, *walk*, *precipitous*, *post*, *walking*, *any*, *antimony*, ***resolute***, *position*, *course*, *spasmodic*, *pointed*, ***obstinate***, *cease*, *thrust*

spr-: ***shoot***, *drops*, *elastic*, *small*, *particles*, ***extend***, *lively*, *germinate*, *breadth*, *alfione*, *picea*, *surffish*, *ungracefully*, *seed*, *sail*, *cause*, *source*, *rhacochilus*, *sharptailed*, *plant*

-Vng: *the*, *art*, *material*, *to*, *business*, ***sound***, *or*, *that*, *collectively*, *boards*, *operation*, *practice*, *from*, *adapted*, *cloth*, *vb*, *etc*, *acid*, *work*, *off*

Detailed results for all 46 phonesthemes evaluated can be found in Appendix A.

6.3.1 Significance Testing

The mutual information method allows a researcher to find a list of definition words that are correlated with a candidate phonestheme's orthographic pattern, sorted by the MI value of the word. It remains to be shown that the definition words selected by these techniques for proposed phonestheme sets are selected at a rate higher than chance—that is, that the form-meaning pairings, in the terms of definition (1), are "better attested in the lexicon of a language than would be predicted, all other things being equal".

One way to test for significance is to compare the results for a candidate phonestheme with those of a randomly-selected set of definitions. If the results for the phonestheme set are more pronounced than for the random set—that is, if MI scores are higher—then the phonestheme is more likely to be real. By repeatedly selecting new random sets and comparing them to the candidate set, it is possible to empirically estimate the *p* value, the likelihood that the result is due to chance.

The precise procedure is as follows. First, create a set of definitions whose headwords match the orthographic pattern of the candidate phonestheme. Next, create a set of definitions that contains every definition in the dictionary. Both sets of definitions have all three filters applied; in particular, both sets are filtered to remove definition words that match the phonesthetic pattern (e.g. every word beginning with *sn*-)—otherwise, words matching the pattern would appear disproportionately often in the non-candidate set. Calculate the mutual information for all definition words using this pair of sets. Next, repeatedly select a random set of definitions with the same number of definitions as the candidate set and calculate the mutual information for that set and the whole dictionary. (In the results reported in Appendix A below, 1000 random sets have been generated for each candidate set to give a good estimate of *p*.) For

each random set, keep track of the MI value for the most informational definition word. Finally, for each definition word in the candidate set, we can estimate the *p* value by comparing its MI value with all 1000 highest MI values for the random sets. If a candidate word's MI is less than the value of the maximum MI values for a random set *n* times, then the empirical estimate of *p* is simply:

(13) $$p = \frac{n}{1000}$$

It is important to note that this first estimate of the *p* value is insensitive to which particular words have occurred with high MI values—I will therefore refer to it as the word-independent *p* value. To see why, consider the results for the phonestheme *cr-* 'harsh or unpleasant noise', in which the definition word *noise* had an estimated *p* value of 0.887, meaning that 887 times out of a thousand, some random word had a higher MI than 0.0000097769. That *p* value is not statistically significant; however, it was calculated without taking account of the identity of the word. The chance that the particular word *noise*, which is clearly related to the meaning of the phonestheme, would occur near the top of the sorted list is very small. Taking account of the meaning of definition words allows us to make a second estimate of significance based on the position of the highest word with a meaning related to the candidate phonestheme in the MI list. If we knew there was only a single definition word that expressed the core meaning of the phonestheme, then assuming that *V* different word types occur in definitions, the chance of that word appearing between positions 1 and *n* (inclusive) on the sorted MI list would be:

(14) $$p = \frac{n}{V}$$

The dictionary used here has 71,459 word types occurring in definitions, so for the case of *cr-* described above, there is only a probability of only about 0.00007 that the word *noise* would occur in the top five. Of course, there are usually multiple definition words that carry the phonesthetic meaning. If there are *w* different words that express the phonestheme's meaning, then the chance of at least one of these appearing between positions 1 and *n* (inclusive) is:

$$p = 1 - \prod_{i=1}^{w} \frac{V - n - i + 1}{V} \tag{15}$$

This formula allows us to calculate the statistical significance of the appearance of definition words associated with the phonesthetic meaning at the top of the sorted MI list. For example, if there were ten definition words with the phonesthetic meaning, the chance of one of them appearing at position 20 or higher is approximately 0.0034, so finding one or more of the them in the top 20 is statistically significant. Unfortunately, calculating this second *p* value is difficult in practice because doing so requires knowing the number of acceptable definition words, but going through the entire 71,459 words for each candidate phonestheme is impractical. Therefore, in the results in Appendix A below I have simply reported the first (word-independent) *p* value and included the top twenty words. For the value $n = 20$, the appearance of a phonesthetically-related word in the list is significant ($p < 0.05$) as long as there are 68 or fewer definition words that express the phonesthetic meaning.

Based on these two tests for significance, the results reported in Appendix A are broken into three groups of candidates. In the first group, labeled "strongly confirmed", the candidate phonestheme has passed both tests—that is, the most highly ranked phonesthetic definition word has a *p* value less than 0.05, and at least one such word appears in the top twenty. In the second group, labeled "weakly confirmed", the word-independent *p* value was not significant, but at

least one phonesthetic word still occurs in the top twenty. In the third group, labeled "unconfirmed", are phonesthemes that passed neither test. Of the 46 phonesthemes tested, four were strongly confirmed, 33 were weakly confirmed, and nine were unconfirmed.

For comparison, after the phonesthemes I have included the results of applying the mutual information method to several etyma and morphemes, including the etyma *-doct-* 'teach', *-viv-* 'life', and *-mit* 'send' and the productive morpheme *un-* 'not'. Intuitively, these results ought to have even stronger form-meaning associations that phonesthemes. This is true for *un-*, but surprisingly not for *-viv-*, *-mit* and *-doct-* (though *-viv-* is close to statistical significance). Of course, the reality of these etyma and morphemes is not controversial, and so these mixed results show only that the MI method is not infallible—non-confirmations just demonstrate a failure of the method, not the non-existence of a form-meaning pairing.

7 Future Work and Conclusion

In the future, these results might be improved by finding another way of scoring definition words that produces even better results than mutual information, or by developing more sophisticated filters that do a better job of remove interfering non-phonestheme words. It would also be interesting to try the MI method using a different dictionary, perhaps one with more consistently worded etymologies. It is also worth noting that, while I have been treating morpheme and etymon detection as false positives, it is possible that the MI method's ability to find them is actually useful. For example, the MI method, used to test the correlation between subword strings of characters and definition words in the lexicon of an understudied language, could be used to produce a set proposed morphemes for that language.

In this paper, I have described the development and evaluation of three statistical methods for detecting and validating phonesthemes that can be applied by a computer. Of these,

the clustering method and the relative word frequency methods failed to produce positive results. The mutual information method, on the other hand, was quite successful. With the addition of the tests for statistical significance, the MI method is even capable of searching for previously unknown phonesthemes by simply applying it, for example, to every attested onset consonant cluster in the target language, then examining the statistically significant definition words for phonesthetic meanings.

8 Acknowledgments

The author wishes to acknowledge the contributions of Paul Sampson, Mark Giganti, and Hilary Lyons of the University of Washington Department of Statistics, who provided invaluable advice on the issue of statistical significance testing.

References

Abelin, Åsa. 1999. Studies in sound symbolism. Göteborg, Sweden: Göteborg University dissertation.

Bergen, Benjamin K. 2004. The psychological reality of phonaesthemes. Language 80.290-311.

Bloomfield, Leonard. 1933. Language. New York: Henry Holt.

Bloomfield, Morton W. 1953. Final root-forming morphemes. American Speech 28.158-164

Bolinger, Dwight L. 1965. Forms of English: Accent, morpheme, order. Cambridge, MA: Harvard University Press.

Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer, Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 No. 1: 391-407.

Fano, Robert M. 1961. Transmission of information: A statistical theory of communications. Cambridge, MA: MIT Press.

Firth, John R. 1930. Speech. In The tongues of men and Speech, ed. by Peter Strevens. Oxford, UK: Oxford University Press.

Hutchins, Sharon Suzanne. 1998. The psychological reality, variability, and compositionality of English phonesthemes. Atlanta, GA: Emory University dissertation.

Jespersen. 1922. Language. London: George Allen & Unwin.

Karypis, George. 2003. CLUTO: Software package for clustering high-dimensional datasets. http://www-users.cs.umn.edu/~karypis/cluto/

Kempley, S. T. and John Morton. 1982. The effects of priming with regularly and irregularly related words in auditory word recognition. British Journal of Psychology 73.441–45.

McCallum, Andrew. 1996. BOW: A toolkit for statistical language modeling, text retrieval,

classification and clustering. http://www-2.cs.cmu.edu/~mccallum/bow/

McCallum, Andrew and Kamal Nigam. 1998. A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization.

Porter, Noah et. al. 1913. Webster's revised unabridged dictionary. ftp://ftp.dict.org/pub/dict/

Rhodes, Richard A. and John M. Lawler. 1981. Athematic metaphors. Papers from the seventeenth regional meeting of the Chicago linguistic society 1981 (CLS 17), ed. by Roberta Hendrick, Carrie Masek, and Mary Frances Miller. 318-342.

Shannon, Claude E. 1948. A mathematical theory of communication. The Bell System Technical Journal 27: 379-423, 623-656.

Wallis, John. 1653. Grammaticae linguae Anglicanae. Menston, UK: Scolar Press.

Westcott, Roger W. 1987. Holesthemes or phonesthemes twice over. General Linguistics 27.67-72.

Appendix A: Detailed Results of the Mutual Information Method

For each phonestheme are listed the orthographic pattern of the phonestheme, a short paraphrase of the meaning tested by Hutchins (1998), the number of headwords in the 1913 Webster's that matched the pattern, and the top twenty definition words with the highest MI scores (with words matching the proposed meaning in boldface). Instead of the value of *p*, I have reported (1 - *p*), so that higher values in this column imply greater statistical significance. The phonesthemes are organized into three groups: STRONGLY CONFIRMED, where the *p* value of the most informational word that has the phonesthetic meaning is less than 0.05; WEAKLY CONFIRMED, in which no single word's *p* value is below 0.05, but one or more words with the phonesthetic meaning do appear in the top 20; and UNCONFIRMED, in which no word with the phonesthetic meaning occurs in the top 20. Also included are the results for the MI method on four non-phonesthemes: the etymon -*viv*- 'life', the etymon -*mit* 'send', the etymon -*doct*- 'teach', and the productive morpheme *un*- 'not'. Phonesthemes are sorted by the most informational word with the proposed meaning, from highest to lowest, except in the case of the unconfirmed phonesthemes, which are sorted alphabetically.

Generally, only words that are synonym or near-synonyms are highlighted, even when words clearly related to the phonesthetic meaning occur. In the lists of definition words, words are not highlighted if they match the phonestheme's orthographic pattern, as sometimes happened for rhyme morphemes (e.g. *pricks* and *nicks* in the list for -*ick*). It is interesting to note that some phonesthemes were confirmed in spite of interference from other words that also fit the pattern (e.g. the list for -*Vng* contains several words associated with the verbal suffix -*ing*). Such interference may have been a factor in the non-confirmation of the phonestheme -*ip*, since it overlaps with the semi-productive morpheme -*ship*.

Strongly Confirmed:

***sn-* 'nose; snobbish' (170)**

def. word	MI	1 - *p*
nose	**0.0000565307**	**0.997**
sharp	0.0000163574	0.673
reprimand	0.0000133541	0.471
seize	0.0000121417	0.332
contempt	**0.0000119126**	**0.312**
short	0.0000118340	0.301
bite	0.0000116533	0.276
with	0.0000097613	0.128
laugh	0.0000097334	0.126
nasal	**0.0000090017**	**0.049**
angry	0.0000088951	0.042
check	0.0000087179	0.034
air	0.0000085600	0.027
nip	0.0000082975	0.017
catch	0.0000082894	0.017
fellow	0.0000082605	0.014
mucus	0.0000081098	0.011
surly	0.0000081098	0.011
rebuke	0.0000079575	0.007
mean	0.0000079168	0.007

***st-* 'firm; upright; linear' (1493)**

def. word	MI	1 - *p*
to	0.0000340000	0.998
firm	**0.0000234677**	**0.975**
fixed	**0.0000201057**	**0.952**
in	0.0000138853	0.749
upright	**0.0000127493**	**0.651**
vessel	0.0000118034	0.548
walk	0.0000104120	0.319
precipitous	0.0000099669	0.257
post	0.0000094312	0.190
walking	0.0000093334	0.177
any	0.0000087957	0.097
antimony	0.0000086452	0.078
resolute	**0.0000085401**	**0.068**
position	0.0000081814	0.044
course	0.0000081642	0.044
spasmodic	0.0000079706	0.032
pointed	0.0000078469	0.028
obstinate	**0.0000077918**	**0.026**
cease	0.0000076854	0.021
thrust	0.0000076060	0.017

***-Vng* 'ringing sound' (2316)**

def. word	MI	1 - *p*
the	0.0000485726	1.000
art	0.0000434058	1.000
material	0.0000314456	0.999
to	0.0000310481	0.999
business	0.0000233477	0.990
sound	**0.0000227960**	**0.988**
or	0.0000221536	0.987
that	0.0000217262	0.985
collectively	0.0000211508	0.984
boards	0.0000204212	0.979
operation	0.0000163196	0.911
practice	0.0000162520	0.907
from	0.0000157789	0.885
adapted	0.0000156229	0.880
cloth	0.0000154758	0.880
vb	0.0000151131	0.869
etc	0.0000125643	0.715
acid	0.0000123925	0.695
work	0.0000121628	0.650
off	0.0000110674	0.534

***spr-* 'to radiate out; elongated' (67)**

def. word	MI	1 - *p*
shoot	**0.0000277869**	**0.951**
drops	0.0000174379	0.797
elastic	0.0000159478	0.716
small	0.0000106687	0.259
particles	0.0000100018	0.176
extend	**0.0000089230**	**0.102**
lively	0.0000085093	0.073
germinate	0.0000082796	0.060
breadth	0.0000072713	0.011
alfione	0.0000069389	0.008
picea	0.0000069389	0.008
surffish	0.0000069389	0.008
ungracefully	0.0000069389	0.008
seed	0.0000069251	0.008
sail	0.0000068950	0.007
cause	0.0000068693	0.006
source	0.0000065946	0.003
rhacochilus	0.0000065616	0.002
sharptailed	0.0000065616	0.002
plant	0.0000064744	0.002

Weakly Confirmed:

***cl-* 'noise from a collision' (468)**

def. word	MI	1 - *p*
together	**0.0000223574**	**0.935**
noise	**0.0000192252**	**0.836**
free	0.0000183885	0.809
fast	0.0000165367	0.730
ringing	**0.0000149044**	**0.630**
collision	**0.0000138590**	**0.531**
sharp	**0.0000130513**	**0.464**
loud	**0.0000115029**	**0.270**
grasp	0.0000113225	0.252
hands	0.0000112173	0.248
striking	**0.0000105062**	**0.186**
with	0.0000091880	0.055
hen	0.0000091780	0.055
noises	**0.0000083983**	**0.020**
rattling	0.0000081774	0.014
hold	0.0000081013	0.013
ascend	0.0000075302	0.003
learned	0.0000073814	0.002
wood	0.0000069803	0.001
embracing	0.0000067705	0.000

***-ash* 'violent action or collision' (76)**

def. word	MI	1 - *p*
sudden	**0.0000234888**	**0.911**
water	0.0000223311	0.898
strike	**0.0000194116**	**0.837**
washed	0.0000188310	0.819
violently	**0.0000162525**	**0.723**
crush	0.0000161565	0.719
whip	0.0000147342	0.619
collision	**0.0000143477**	**0.581**
break	**0.0000141867**	**0.566**
dashing	0.0000121185	0.395
pieces	0.0000121067	0.393
noise	0.0000116438	0.355
of	0.0000115997	0.353
cut	0.0000112792	0.326
burst	0.0000107014	0.237
potassium	0.0000091930	0.096
ashes	0.0000090395	0.084
noisily	0.0000089725	0.082
random	0.0000087781	0.074
ablution	0.0000087564	0.072

***sp-* 'send out; reject' (917)**

def. word	MI	1 - *p*
small	0.0000213934	0.934
shoot	**0.0000206637**	**0.928**
slender	0.0000143688	0.657
semen	0.0000130940	0.538
saliva	0.0000118240	0.402
lively	0.0000116419	0.385
scattered	0.0000102828	0.200
emit	**0.0000102115**	**0.190**
long	0.0000095098	0.112
jet	**0.0000092682**	**0.091**
out	**0.0000089346**	**0.063**
eject	**0.0000079452**	**0.016**
thorny	0.0000079254	0.015
drops	0.0000078084	0.009
elastic	0.0000075880	0.005
apparition	0.0000073138	0.001
pintail	0.0000069990	0.000
occuring	0.0000068887	0.000
sail	0.0000068624	0.000
seminal	0.0000068254	0.000

***sl-* 'slide; careless' (316)**

def. word	MI	1 - *p*
snow	0.0000325466	0.980
smooth	**0.0000231959**	**0.903**
cut	0.0000222337	0.876
lazy	**0.0000155114**	**0.631**
ice	0.0000154357	0.628
runners	0.0000145337	0.557
oblique	0.0000127211	0.365
narrow	0.0000122837	0.325
not	0.0000121585	0.313
imp	0.0000114737	0.250
loose	0.0000113995	0.238
negligent	**0.0000112742**	**0.228**
carelessly	**0.0000111620**	**0.218**
weavers	0.0000106428	0.159
prov	0.0000105505	0.148
saliva	0.0000104607	0.138
eng	0.0000103858	0.134
readymade	0.0000100859	0.110
spill	0.0000100859	0.110
smoothly	**0.0000099513**	**0.103**

Weakly Confirmed (continued):

***-ick* 'sudden; abrupt; sharp' (97)**

def. word	MI	1 - *p*
pointed	**0.0000214820**	**0.871**
sharp	**0.0000185220**	**0.792**
strike	0.0000147027	0.608
attach	0.0000145464	0.596
nicks	0.0000124814	0.419
pricks	0.0000109068	0.278
with	0.0000105160	0.219
backsword	0.0000093607	0.098
thrust	0.0000084413	0.044
mark	0.0000081799	0.036
point	**0.0000080698**	**0.033**
tongue	0.0000080078	0.031
notch	0.0000076993	0.022
hit	0.0000072864	0.010
puncturing	**0.0000072374**	**0.010**
up	0.0000071286	0.007
dog	0.0000070878	0.007
puncture	**0.0000069746**	**0.004**
ticks	0.0000069746	0.004
picking	0.0000068558	0.001

***gl-* 'light; vision' (365)**

def. word	MI	1 - *p*
smooth	0.0000232839	0.913
specious	0.0000222555	0.894
spherical	0.0000200744	0.840
look	**0.0000186537**	**0.802**
sullen	0.0000183769	0.795
light	**0.0000181011**	**0.784**
shine	**0.0000179517**	**0.778**
viscous	0.0000157358	0.678
bright	**0.0000121656**	**0.356**
luster	0.0000120111	0.343
ice	0.0000116167	0.310
stare	**0.0000114393**	**0.292**
acid	0.0000114003	0.290
comments	0.0000106663	0.210
sugar	0.0000101909	0.152
white	0.0000100298	0.134
and	0.0000088907	0.049
dilute	0.0000088024	0.042
vitreous	0.0000088024	0.042
commentator	0.0000086735	0.040

***-olt* 'energetic force in motion' (36)**

def. word	MI	1 - *p*
electromotive	**0.0000167597**	**0.810**
bolts	0.0000125978	0.568
arrow	0.0000123170	0.539
coupling	0.0000121825	0.521
revolts	0.0000110737	0.396
jolts	0.0000107473	0.362
nomination	0.0000094853	0.243
party	0.0000091003	0.209
sudden	**0.0000090787**	**0.208**
spring	**0.0000088445**	**0.182**
pin	0.0000082582	0.106
lightning	0.0000075468	0.035
caucus	0.0000072683	0.028
shake	**0.0000070212**	**0.019**
bolter	0.0000069790	0.019
shock	**0.0000069396**	**0.018**
suddenly	**0.0000068056**	**0.012**
hagdon	0.0000067441	0.011
smites	0.0000067441	0.011
voussoirs	0.0000067441	0.011

***fl-* 'motion, repeated or fluid' (573)**

def. word	MI	1 - *p*
light	0.0000203956	0.879
surface	0.0000183926	0.813
move	**0.0000180981**	**0.801**
sudden	0.0000172707	0.775
with	0.0000160018	0.710
to	0.0000138013	0.552
throw	0.0000131845	0.502
wings	0.0000130130	0.482
burst	0.0000125121	0.423
fan	**0.0000118114**	**0.331**
level	0.0000113020	0.242
air	**0.0000104473**	**0.162**
broad	0.0000096898	0.104
water	**0.0000095011**	**0.083**
ebb	**0.0000091774**	**0.055**
side	0.0000089909	0.043
stream	**0.0000089771**	**0.043**
glass	0.0000088981	0.037
loose	0.0000086098	0.025
pitch	0.0000085989	0.025

Weakly Confirmed (continued):

***scr-/skr-* 'sound; irregular mov.' (151)**

def. word	MI	1 - *p*
writing	0.0000251849	0.929
rub	0.0000185727	0.816
shrill	**0.0000180223**	**0.796**
stunted	0.0000147392	0.607
of	0.0000138655	0.549
rough	0.0000119504	0.335
shriek	**0.0000108365**	**0.223**
hastily	0.0000098961	0.138
irregular	**0.0000094197**	**0.084**
lean	0.0000089167	0.045
brush	0.0000085242	0.028
struggle	0.0000083892	0.025
something	0.0000083478	0.024
rubbing	0.0000078657	0.015
sharp	0.0000078228	0.015
writer	0.0000076775	0.013
drawing	0.0000074732	0.009
across	0.0000074607	0.009
examination	0.0000074189	0.009
fours	0.0000071851	0.008

***-inge* 'spasm; contraction; pain' (27)**

def. word	MI	1 - *p*
contract	**0.0000149173**	**0.742**
burn	0.0000100430	0.329
hinges	0.0000098040	0.313
constrict	**0.0000074668**	**0.040**
tweak	0.0000072317	0.035
peristome	0.0000070336	0.030
servility	0.0000064552	0.016
pinch	**0.0000063442**	**0.012**
transgress	0.0000061476	0.007
sudden	**0.0000060009**	**0.006**
sharp	**0.0000058882**	**0.006**
border	0.0000057972	0.004
darting	0.0000057587	0.004
interference	0.0000056314	0.004
lash	0.0000053600	0.003
compress	**0.0000053120**	**0.003**
depend	0.0000052213	0.003
cardinal	0.0000051783	0.003
together	0.0000050957	0.003
shrink	**0.0000050575**	**0.003**

***sw-* 'move rhythmically' (251)**

def. word	MI	1 - *p*
motion	**0.0000179022**	**0.795**
broom	0.0000121259	0.367
imp	0.0000119097	0.347
tawny	0.0000117420	0.320
oath	0.0000117349	0.319
cleaning	0.0000104282	0.162
drink	0.0000103597	0.155
sink	0.0000098151	0.110
with	0.0000087617	0.031
bully	0.0000085382	0.019
hogsty	0.0000082441	0.016
perspire	0.0000078504	0.006
long	0.0000077051	0.003
clean	0.0000074085	0.002
move	**0.0000073560**	**0.001**
winning	0.0000072525	0.001
toil	0.0000072472	0.001
brush	0.0000068104	0.001
brushing	0.0000067982	0.001
singe	0.0000067982	0.001

***-irl/-url* 'twist; intertwine' (31)**

def. word	MI	1 - *p*
curls	0.0000195952	0.879
whirling	0.0000189208	0.863
twist	**0.0000156504**	**0.741**
revolve	**0.0000136310**	**0.641**
eddy	**0.0000136074**	**0.641**
hurling	0.0000126944	0.561
ringlets	0.0000103971	0.319
rapidly	0.0000103721	0.316
velocity	0.0000096200	0.253
undulations	0.0000090175	0.192
obstructions	0.0000088309	0.160
curled	0.0000080427	0.068
hair	0.0000074679	0.032
with	0.0000072763	0.022
motion	**0.0000069689**	**0.016**
spirals	**0.0000066310**	**0.010**
move	**0.0000065440**	**0.009**
crossgrained	0.0000064803	0.007
the	0.0000064472	0.007
beer	0.0000063931	0.006

Weakly Confirmed (continued):

***tw-* 'turn; distort' (113)**

def. word	MI	1 - *p*
winding	**0.0000171478**	**0.723**
nineteen	0.0000138728	0.519
units	0.0000134105	0.471
next	0.0000134076	0.471
intermitted	0.0000126831	0.392
pull	0.0000118677	0.308
convolution	**0.0000116353**	**0.274**
pinch	0.0000114298	0.258
after	0.0000113813	0.254
parts	0.0000108282	0.216
divided	0.0000100825	0.134
quick	0.0000099960	0.125
gabble	0.0000096822	0.097
spirally	**0.0000096619**	**0.097**
birth	0.0000091355	0.066
jerk	0.0000091355	0.066
torsion	**0.0000090781**	**0.062**
one	0.0000089616	0.054
wreathe	0.0000086219	0.045
wink	0.0000084294	0.039

***wr-* 'irregular motion; twist' (112)**

def. word	MI	1 - *p*
distorted	**0.0000146704**	**0.637**
distort	**0.0000145923**	**0.627**
twisted	**0.0000125965**	**0.450**
angry	0.0000125146	0.440
violence	0.0000112813	0.336
ruin	0.0000112466	0.334
shipwreck	0.0000111971	0.326
pervert	0.0000097421	0.125
characters	0.0000095581	0.103
twisting	**0.0000082798**	**0.031**
involve	0.0000081598	0.023
anger	0.0000080937	0.019
extort	0.0000077738	0.015
turn	**0.0000076071**	**0.013**
unjustly	0.0000071729	0.005
twist	**0.0000068691**	**0.000**
dispute	0.0000068347	0.000
right	0.0000068213	0.000
miserable	0.0000068009	0.000
as	0.0000066521	0.000

***sc-/sk-* 'surface; edge; thin' (938)**

def. word	MI	1 - *p*
induration	0.0000154477	0.725
surface	**0.0000148562**	**0.692**
rough	0.0000132500	0.544
coat	0.0000122905	0.442
cut	0.0000113386	0.339
thin	**0.0000112649**	**0.324**
writing	0.0000109730	0.286
rub	0.0000108629	0.271
brush	0.0000107780	0.258
bony	0.0000107456	0.252
superficially	**0.0000103004**	**0.197**
shrill	0.0000102949	0.196
run	0.0000102234	0.184
knowledge	0.0000096637	0.110
hastily	0.0000096621	0.110
edge	**0.0000094887**	**0.090**
small	0.0000094263	0.088
stunted	0.0000093918	0.085
mark	0.0000091862	0.073
struggle	0.0000089594	0.061

***-awl* 'slow; stretched' (22)**

def. word	MI	1 - *p*
slow	**0.0000128420**	**0.614**
cry	0.0000116755	0.504
loud	0.0000114361	0.489
spittle	0.0000099487	0.371
creeping	**0.0000085998**	**0.190**
ungracefully	0.0000085241	0.178
waul	0.0000085241	0.178
saddlers	0.0000081461	0.130
inelegantly	0.0000076214	0.087
unskillfully	0.0000072520	0.063
slowly	**0.0000071979**	**0.061**
ratchet	0.0000071011	0.054
limbs	0.0000070015	0.050
lengthened	**0.0000069664**	**0.049**
scribble	0.0000069664	0.049
shoemakers	0.0000069664	0.049
advance	0.0000062036	0.023
creep	**0.0000061473**	**0.019**
move	0.0000058819	0.016
spread	0.0000057080	0.014

Weakly Confirmed (continued):

***str-* 'linear; forceful action' (337)**

def. word	MI	1 - *p*
narrow	**0.0000145430**	**0.567**
wander	0.0000126039	0.363
force	**0.0000121471**	**0.317**
effort	**0.0000098820**	**0.084**
ostriches	0.0000097624	0.082
blow	0.0000097241	0.079
extend	**0.0000093615**	**0.056**
shrill	0.0000091543	0.053
efforts	**0.0000090490**	**0.049**
instrument	0.0000089795	0.048
variant	0.0000083508	0.020
line	**0.0000078391**	**0.005**
piston	0.0000075273	0.001
apart	0.0000074958	0.001
layers	0.0000073124	0.000
course	0.0000071581	0.000
clock	0.0000071525	0.000
movement	**0.0000069809**	**0.000**
conch	0.0000069075	0.000
rigorously	**0.0000069075**	**0.000**

***-isp* 'swift or bounded motion' (5)**

def. word	MI	1 - *p*
brittle	0.0000136610	0.812
ripple	0.0000119844	0.754
fatuus	0.0000093642	0.504
ignis	0.0000090744	0.482
undulate	**0.0000086408**	**0.455**
ringlets	0.0000083184	0.422
crackling	0.0000079502	0.384
speak	0.0000077732	0.367
pronounce	0.0000061298	0.226
articulation	0.0000059192	0.167
imperfectly	0.0000054853	0.050
imperfect	0.0000052615	0.022
lively	0.0000050281	0.011
childlike	0.0000048695	0.005
mispronounce	0.0000048695	0.005
sparking	0.0000048695	0.005
unwilted	0.0000048695	0.005
with	0.0000047304	0.004
hesitatingly	0.0000045355	0.000
express	0.0000042803	0.000

***tr-* 'path; line; go on foot' (1237)**

def. word	MI	1 - *p*
three	0.0002829026	1.000
another	0.0000461530	1.000
change	0.0000214925	0.960
threefold	0.0000181348	0.902
victory	0.0000174582	0.883
barter	0.0000156309	0.804
into	0.0000154139	0.786
conveyance	0.0000137266	0.643
one	0.0000136525	0.641
through	0.0000135893	0.637
foot	**0.0000121923**	**0.494**
goods	0.0000112620	0.393
exchange	0.0000103491	0.284
angles	0.0000103254	0.283
pass	0.0000100619	0.241
third	0.0000092330	0.112
each	0.0000090558	0.094
passing	0.0000089021	0.078
commodities	0.0000083704	0.056
journey	**0.0000081694**	**0.050**

***-ump* 'heavy; low; compact' (34)**

def. word	MI	1 - *p*
plunger	0.0000165042	0.788
plumper	0.0000143198	0.688
card	0.0000131850	0.633
water	0.0000130383	0.622
heavy	**0.0000113136**	**0.412**
stub	0.0000097396	0.281
piston	0.0000095810	0.271
stumps	0.0000094978	0.264
lifts	0.0000090911	0.233
protuberance	0.0000087402	0.195
piece	0.0000085151	0.156
bittern	0.0000083451	0.119
leap	0.0000083451	0.119
jumping	0.0000080074	0.080
considerable	0.0000078009	0.068
blow	0.0000077279	0.060
delivering	0.0000072529	0.029
brokenly	0.0000067817	0.019
bodice	0.0000064926	0.015
heavily	**0.0000063412**	**0.011**

Weakly Confirmed (continued):

***bl-* 'blow; swell; inflate' (446)**

def. word	MI	1 - *p*
color	0.0000281339	0.977
eyes	0.0000145535	0.608
stain	0.0000138047	0.550
happiness	0.0000134803	0.519
air	**0.0000118927**	**0.311**
noisy	0.0000117701	0.303
dim	0.0000111395	0.237
ink	0.0000093418	0.067
sight	0.0000091295	0.054
stupid	0.0000089821	0.043
whiten	0.0000088591	0.037
flowers	0.0000086613	0.023
turgid	**0.0000085546**	**0.018**
make	0.0000085155	0.016
scurrilous	0.0000084132	0.015
censure	0.0000080494	0.010
sap	0.0000079744	0.010
fish	0.0000078772	0.008
paper	0.0000078001	0.005
shedding	0.0000077917	0.005

***-oop* 'curved; concave' (25)**

def. word	MI	1 - *p*
cough	0.0000158464	0.776
hoops	0.0000158464	0.776
whooping	0.0000136837	0.684
forward	0.0000114078	0.469
downward	0.0000100373	0.346
bend	**0.0000093572**	**0.288**
cry	0.0000091671	0.270
prey	0.0000082385	0.127
dipping	**0.0000078425**	**0.083**
centerboard	0.0000075075	0.056
drooped	0.0000075075	0.056
shoveling	0.0000075075	0.056
deck	0.0000073326	0.041
hoot	0.0000071299	0.036
stooping	0.0000071299	0.036
hoopoe	0.0000068407	0.028
tubs	0.0000064080	0.018
halloo	0.0000062370	0.013
tippet	0.0000062370	0.013
barrel	0.0000059618	0.009

***dr-* 'pulling down; languid' ()**

def. word	MI	1 - *p*
water	0.0000203797	0.841
fall	0.0000196189	0.827
along	0.0000190488	0.807
moisture	0.0000164043	0.679
let	0.0000158297	0.650
coupling	0.0000143286	0.534
rain	0.0000119417	0.316
pulling	**0.0000116582**	**0.299**
onward	0.0000106808	0.188
wet	0.0000105955	0.171
liquors	0.0000104747	0.147
slowly	**0.0000102183**	**0.114**
trickling	0.0000100316	0.110
liquid	0.0000097973	0.086
trail	0.0000090323	0.047
tragacanth	0.0000089801	0.047
link	0.0000087663	0.042
lees	0.0000085218	0.033
depth	0.0000084529	0.032
heavy	**0.0000079752**	**0.009**

***-amp* 'restrain; force into a space' (31)**

def. word	MI	1 - *p*
foot	0.0000154327	0.761
incandescent	0.0000129458	0.629
huts	0.0000121499	0.541
tents	0.0000109508	0.406
stamped	0.0000103003	0.335
forcibly	**0.0000095304**	**0.261**
wick	0.0000089097	0.192
sink	0.0000076956	0.048
capsize	0.0000072613	0.022
carbonic	0.0000071471	0.020
aphlogistic	0.0000068839	0.017
imprinted	0.0000068839	0.017
mark	0.0000068541	0.017
crush	**0.0000068164**	**0.015**
boot	0.0000067180	0.015
impress	0.0000066707	0.014
lumbermen	0.0000063601	0.009
wet	0.0000063393	0.008
bite	0.0000061447	0.007
humid	0.0000058410	0.005

Weakly Confirmed (continued):

-*Vnk* 'sharp movement w/ sound' (130)

def. word	MI	1 - *p*
of	0.0000119253	0.328
tinder	0.0000110850	0.250
sharp	**0.0000104540**	**0.200**
mound	0.0000095812	0.096
ranks	0.0000088378	0.035
piece	0.0000080405	0.020
void	0.0000079381	0.018
calf	0.0000077407	0.013
who	0.0000076863	0.013
aimed	0.0000076461	0.013
eyelids	0.0000076461	0.013
drawbar	0.0000076219	0.012
connecting	0.0000075146	0.011
sonorous	**0.0000072211**	**0.009**
postage	0.0000069775	0.003
screw	0.0000069062	0.002
tinkling	0.0000066458	0.000
hole	0.0000065368	0.000
imbibe	0.0000064998	0.000
banker	0.0000061201	0.000

***spl*- 'diverge; spread from a point' (72)**

def. word	MI	1 - *p*
viscera	0.0000142250	0.587
fretful	0.0000112658	0.335
piece	0.0000110257	0.302
bone	0.0000104176	0.204
incision	0.0000099240	0.148
divide	**0.0000092334**	**0.104**
spatter	**0.0000092248**	**0.104**
player	0.0000084101	0.059
dealt	0.0000080282	0.048
mud	0.0000078993	0.045
into	0.0000078989	0.045
two	0.0000077526	0.042
thin	0.0000072078	0.007
blackjack	0.0000066984	0.002
melancholy	0.0000065227	0.002
dash	0.0000064070	0.002
affected	0.0000062104	0.001
anatomy	0.0000061619	0.001
visceral	0.0000060949	0.000
broken	0.0000056996	0.000

***cr*- 'harsh or unpleasant noise' (750)**

def. word	MI	1 - *p*
across	0.0000174963	0.824
iron	0.0000144700	0.660
brittle	0.0000125792	0.478
lame	0.0000113637	0.318
noise	**0.0000097769**	**0.113**
undigested	0.0000093014	0.077
broken	0.0000088534	0.046
with	0.0000087838	0.043
polychroite	0.0000085371	0.031
cipher	0.0000081701	0.014
wrinkles	0.0000081501	0.014
athwart	0.0000078155	0.007
ringlets	0.0000072306	0.003
belief	0.0000072119	0.003
to	0.0000070042	0.002
reptile	0.0000069530	0.002
wrinkle	0.0000069530	0.002
low	0.0000069086	0.002
bar	0.0000068681	0.002
confidence	0.0000068667	0.002

***gr*- 'deep or complaining noise' (609)**

def. word	MI	1 - *p*
steps	0.0000136641	0.552
hard	0.0000130478	0.478
step	0.0000113664	0.277
color	0.0000109480	0.219
etc	0.0000095461	0.103
harsh	0.0000094029	0.088
degrees	0.0000091818	0.061
surly	**0.0000089255**	**0.040**
to	0.0000088489	0.035
clutch	0.0000088421	0.035
herbage	0.0000088421	0.035
sorrow	0.0000087141	0.028
particles	0.0000086929	0.028
wheat	0.0000084120	0.025
aud	0.0000081915	0.020
deep	**0.0000080379**	**0.018**
tend	0.0000080278	0.017
sandstone	0.0000068329	0.000
seizure	0.0000068329	0.000
mercy	0.0000067433	0.000

Weakly Confirmed (continued):

***sp_t* 'a rush of liquid' (81)**

def. word	MI	1 - *p*
jet	**0.0000085446**	**0.040**
alfione	0.0000081240	0.021
nasals	0.0000081240	0.021
surffish	0.0000081240	0.021
woodpecker	0.0000077594	0.006
encasement	0.0000077034	0.006
rhacochilus	0.0000077034	0.006
semivowels	0.0000077034	0.006
splints	0.0000077034	0.006
out	0.0000075460	0.003
spectroscope	0.0000073811	0.002
small	0.0000073241	0.001
toxotes	0.0000071195	0.000
germinate	0.0000068991	0.000
devotes	0.0000063912	0.000
shoot	**0.0000063161**	**0.000**
breathing	0.0000062289	0.000
emergency	0.0000061324	0.000
cleave	0.0000059140	0.000
mockery	0.0000056393	0.000

***-oil* 'liquids or cooking' (65)**

def. word	MI	1 - *p*
boiling	0.0000266432	0.945
foils	0.0000164013	0.721
foliation	0.0000161066	0.710
to	0.0000157169	0.698
of	0.0000130409	0.507
plunder	0.0000107153	0.252
clover	0.0000102264	0.189
boils	0.0000090451	0.093
confusion	0.0000089855	0.091
ornamental	0.0000075485	0.020
defile	0.0000074833	0.013
heat	**0.0000072698**	**0.005**
pillage	0.0000070155	0.003
cylindrically	0.0000070001	0.003
tormentil	0.0000070001	0.003
toils	0.0000066228	0.001
commotion	0.0000065857	0.001
olive	0.0000061917	0.000
medic	0.0000060991	0.000
divisions	0.0000060027	0.000

***-owl* 'sinister thing or action' (40)**

def. word	MI	1 - *p*
cry	0.0000161367	0.769
mournful	0.0000113375	0.393
dog	0.0000097014	0.228
auk	0.0000096661	0.227
ball	0.0000088284	0.163
sound	0.0000087810	0.161
utter	0.0000084975	0.128
domestic	0.0000078145	0.043
threatening	**0.0000076961**	**0.036**
brows	0.0000071397	0.012
look	0.0000068181	0.006
bird	0.0000066309	0.005
bowled	0.0000066198	0.005
frown	0.0000066198	0.005
frowning	0.0000064851	0.004
prey	0.0000064366	0.004
wail	0.0000062523	0.003
bowls	0.0000061502	0.003
grumbling	0.0000061502	0.003
owls	0.0000061502	0.003

***-ack* 'collision; noise; abrupt' (155)**

def. word	MI	1 - *p*
to	0.0000208922	0.858
larch	0.0000099455	0.147
bug	0.0000095055	0.092
pile	0.0000088555	0.037
ridge	0.0000088555	0.037
backward	0.0000087952	0.035
buss	0.0000084535	0.022
hire	0.0000080978	0.012
barracks	0.0000080594	0.012
dowitcher	0.0000080594	0.012
eng	0.0000079863	0.011
hay	0.0000078030	0.010
rear	0.0000078006	0.010
frame	0.0000076906	0.010
cabbage	0.0000072527	0.004
alewife	0.0000070059	0.003
remiss	0.0000070059	0.003
noises	**0.0000068150**	**0.003**
flaw	0.0000066419	0.002
packs	0.0000064836	0.002

Weakly Confirmed (continued):

***squ-* 'soft; spongy; compressed' (121)**

def. word	**MI**	1 - ***p***
scales	0.0000218435	0.890
angles	0.0000129317	0.475
bone	0.0000123858	0.419
cry	0.0000111341	0.294
axes	0.0000109889	0.280
obliquely	0.0000103707	0.222
right	0.0000095408	0.111
hams	0.0000094869	0.107
shrill	0.0000091942	0.080
quinsy	0.0000091613	0.077
scream	0.0000082367	0.031
coincident	0.0000079031	0.018
of	0.0000074413	0.006
temporal	0.0000074325	0.006
heels	0.0000069611	0.004
plump	**0.0000067845**	**0.001**
soft	**0.0000067825**	**0.001**
correspondending	0.0000065879	0.000
crosseyed	0.0000065879	0.000
mutans	0.0000065879	0.000

Unconfirmed:

***-am* 'restrain in a small space' (189)**

def. word	MI	1 - *p*
gong	0.0000112075	0.261
streams	0.0000108606	0.215
who	0.0000084748	0.023
hydraulic	0.0000081699	0.014
froth	0.0000076759	0.006
freak	0.0000059953	0.000
light	0.0000059516	0.000
lever	0.0000058684	0.000
carpinus	0.0000058057	0.000
memorizing	0.0000058057	0.000
slams	0.0000058057	0.000
solidissima	0.0000058057	0.000
spisula	0.0000058057	0.000
streamed	0.0000058057	0.000
occupy	0.0000057051	0.000
tracing	0.0000054674	0.000
clangor	0.0000054299	0.000
madhouse	0.0000054299	0.000
pagellus	0.0000054299	0.000
reprisal	0.0000054299	0.000

***-asp* 'harsh or grating noise' (17)**

def. word	MI	1 - *p*
file	0.0000127384	0.626
embrace	0.0000118376	0.549
hold	0.0000110415	0.487
arms	0.0000106243	0.459
breath	0.0000098695	0.401
haje	0.0000081857	0.160
rasps	0.0000081857	0.160
convulsively	0.0000078961	0.138
pant	0.0000076609	0.114
clasping	0.0000071407	0.081
shut	0.0000071022	0.076
fasten	0.0000064769	0.043
catch	0.0000063310	0.036
staple	0.0000061214	0.030
comprehend	0.0000058374	0.025
grasping	0.0000057875	0.025
seizure	0.0000055235	0.024
with	0.0000048651	0.004
respiration	0.0000047693	0.002
catching	0.0000047264	0.001

***-ap* 'bounded thing or action' (110)**

def. word	MI	1 - *p*
whaup	0.0000261398	0.933
of	0.0000225702	0.888
blow	0.0000135862	0.504
catch	0.0000133384	0.482
strike	0.0000130879	0.454
sharp	0.0000128818	0.430
who	0.0000117445	0.317
involve	0.0000116248	0.307
snaps	0.0000113996	0.293
crack	0.0000112979	0.282
laps	0.0000110523	0.261
quick	0.0000107343	0.247
liquor	0.0000106641	0.242
insnare	0.0000099702	0.140
sudden	0.0000099461	0.139
together	0.0000097435	0.120
with	0.0000097265	0.118
cover	0.0000095578	0.096
broad	0.0000093947	0.076
something	0.0000084801	0.035

***-ip* 'quick movement or action' (417)**

def. word	MI	1 - *p*
office	0.0003148235	1.000
of	0.0000398346	0.986
dignity	0.0000289241	0.976
skill	0.0000233392	0.925
the	0.0000223715	0.906
position	0.0000210027	0.872
personality	0.0000206979	0.858
condition	0.0000164920	0.695
being	0.0000160896	0.680
slips	0.0000150358	0.619
off	0.0000149070	0.609
lash	0.0000116416	0.266
footing	0.0000106075	0.170
rank	0.0000105299	0.163
cutting	0.0000105119	0.163
character	0.0000103407	0.142
lips	0.0000098893	0.084
board	0.0000092905	0.049
tear	0.0000086817	0.019
vessel	0.0000086547	0.018

Unconfirmed (continued):

***-ouch* 'careless; slovenly; low' (22)**

def. word	MI	1 - *p*
of	0.0000228191	0.948
bed	0.0000114106	0.493
touchstone	0.0000107790	0.438
tactile	0.0000096770	0.338
stoop	0.0000095169	0.317
side	0.0000083761	0.155
slight	0.0000080880	0.122
contact	0.0000079742	0.106
affect	0.0000074481	0.064
repose	0.0000073697	0.062
emerges	0.0000070001	0.044
warrant	0.0000068548	0.043
escutcheon	0.0000067455	0.040
on	0.0000066870	0.037
darkly	0.0000059894	0.016
jewel	0.0000057765	0.013
down	0.0000055123	0.012
attestation	0.0000055123	0.012
chevron	0.0000054352	0.012
fess	0.0000053626	0.012

***str_p* 'line having breadth' (3)**

def. word	MI	1 - *p*
razor	0.0000088906	0.604
sharpen	0.0000087161	0.593
shoulder	0.0000066324	0.392
spliced	0.0000051176	0.022
deprive	0.0000047633	0.004
bereave	0.0000045994	0.002
rifled	0.0000044901	0.001
chastise	0.0000040629	0.000
projectile	0.0000038071	0.000
peel	0.0000037727	0.000
acquiring	0.0000037091	0.000
farrow	0.0000037091	0.000
trough	0.0000036510	0.000
pliable	0.0000035486	0.000
wreath	0.0000035486	0.000
specifically	0.0000034405	0.000
issuing	0.0000033126	0.000
sheath	0.0000031789	0.000
exclusive	0.0000030458	0.000
grasses	0.0000030243	0.000

***sm-* 'insulting, pejorative term' (140)**

def. word	MI	1 - *p*
spruce	0.0000150722	0.624
blacken	0.0000143101	0.578
slight	0.0000126282	0.419
tobacco	0.0000112366	0.275
pungent	0.0000110003	0.256
soil	0.0000104003	0.204
stain	0.0000103710	0.204
soot	0.0000100020	0.170
merganser	0.0000097474	0.129
buss	0.0000097340	0.126
ustilago	0.0000093379	0.087
olfactory	0.0000093046	0.079
scent	0.0000086937	0.047
superficial	0.0000085408	0.038
sebaceous	0.0000077516	0.015
emerald	0.0000076044	0.013
export	0.0000076044	0.013
quick	0.0000073536	0.011
dirty	0.0000070180	0.004
frock	0.0000070036	0.003

***-ust* 'formation on a surface' (58)**

def. word	MI	1 - *p*
reliance	0.0000157168	0.707
incrusted	0.0000114705	0.395
confidence	0.0000103835	0.243
credit	0.0000098780	0.192
confide	0.0000087947	0.101
incrustation	0.0000078569	0.051
push	0.0000074317	0.018
hope	0.0000071254	0.006
musty	0.0000069853	0.005
suspicion	0.0000063369	0.001
mustiness	0.0000061668	0.001
reposed	0.0000059321	0.001
lists	0.0000055638	0.001
scorched	0.0000055638	0.001
future	0.0000055102	0.001
confidently	0.0000054135	0.000
grasshoppers	0.0000054135	0.000
mildew	0.0000054135	0.000
distaste	0.0000051580	0.000
sell	0.0000050489	0.000

Unconfirmed (continued):

-*Vsk* ‘brief movement or action’ (192)

def. word	**MI**	1 - ***p***
boscage	0.0000124626	0.397
pinefinch	0.0000124626	0.397
disguise	0.0000111844	0.237
sweeping	0.0000083395	0.015
spinus	0.0000078946	0.006
gayety	0.0000072138	0.004
caper	0.0000068416	0.003
conceal	0.0000064854	0.003
skip	0.0000064303	0.003
argophylla	0.0000062309	0.003
eurybia	0.0000062309	0.003
frolicsome	0.0000058760	0.002
casque	0.0000058543	0.002
gambol	0.0000058543	0.002
torsk	0.0000058543	0.002
covering	0.0000057120	0.002
wapiti	0.0000053320	0.002
cover	0.0000053248	0.002
lazy	0.0000052943	0.002
banns	0.0000051351	0.002

Etyma and Morphemes (continued):

***-doct-* 'teach' (21)**

def. word	MI	1 - *p*
physician	0.0000112900	0.477
principles	0.0000088772	0.278
teaching	**0.0000083944**	**0.212**
teach	**0.0000080583**	**0.158**
hydropathist	0.0000077334	0.122
learning	**0.0000064144**	**0.034**
diseases	0.0000060696	0.023
imbue	0.0000055057	0.014
degree	0.0000052894	0.013
rudiments	0.0000052387	0.013
confer	0.0000051100	0.011
teacher	**0.0000051100**	**0.011**
instruct	**0.0000047698**	**0.008**
branch	0.0000046311	0.004
title	0.0000046108	0.004
learned	**0.0000044549**	**0.003**
taught	**0.0000043720**	**0.002**
calicoprinting	0.0000040380	0.000
profession	0.0000039941	0.000
instruction	**0.0000038438**	**0.000**

***un-* 'not' (1778)**

def. word	MI	1 - *p*
not	**0.0008417967**	**1.000**
to	0.0001305533	1.000
deprive	0.0001015683	1.000
remove	0.0000838100	1.000
from	0.0000624042	1.000
loose	0.0000556002	1.000
no	**0.0000472605**	**1.000**
free	0.0000403630	1.000
divest	0.0000401575	1.000
take	0.0000399096	1.000
and	0.0000397991	1.000
the	0.0000384430	1.000
open	0.0000353778	1.000
want	0.0000282414	0.992
subordinate	0.0000256939	0.972
strip	0.0000223010	0.951
release	0.0000208960	0.928
absence	**0.0000201421**	**0.909**
which	0.0000200108	0.905
beneath	0.0000199441	0.901

***-mit* 'send' (33)**

def. word	MI	1 - *p*
to	0.0000704872	0.997
send	**0.0000189852**	**0.787**
leave	0.0000134202	0.568
resign	0.0000108411	0.341
give	**0.0000102946**	**0.286**
refer	0.0000094794	0.213
eject	0.0000089135	0.168
yield	0.0000086222	0.153
emits	0.0000083137	0.134
allow	0.0000080480	0.111
of	0.0000078313	0.084
pass	0.0000069403	0.022
puke	0.0000068355	0.021
abate	0.0000067877	0.019
remits	0.0000064919	0.008
limits	0.0000061058	0.005
admitted	0.0000059599	0.003
permission	0.0000058642	0.003
spew	0.0000058347	0.003
license	0.0000058033	0.003

***-viv-* 'life' (70)**

def. word	MI	1 - *p*
life	**0.0000302398**	**0.931**
alive	**0.0000160519**	**0.596**
renewed	0.0000124011	0.365
lively	**0.0000119915**	**0.326**
live	**0.0000107097**	**0.225**
recover	0.0000087278	0.062
living	**0.0000082730**	**0.042**
festivity	0.0000081149	0.038
interest	0.0000071661	0.011
restoration	0.0000067135	0.003
metal	0.0000066876	0.003
animate	**0.0000064801**	**0.000**
outlive	0.0000064621	0.000
oviparous	0.0000060681	0.000
houseleek	0.0000057753	0.000
restore	0.0000057280	0.000
metallic	0.0000055132	0.000
feast	0.0000053711	0.000
depression	0.0000052971	0.000
joint	0.0000051285	0.000