Statistical Techniques for Detecting and Validating Phonesthemes

Scott Drellishak
University of Washington
sfd@u.washington.edu

- **Phonesthemes**
- Psycholinguistic experiments
- Statistical methods
- Procedure and results
- Closing Remarks

Phonesthemes

- Consider these sound-meaning patterns in the lexicon of English:

  gl- is associated with light or vision:
  glisten, *glitter*, *gleam*, *glow*, *glint*, …

  sn- is associated with the nose:
  sniff, *sneeze*, *snout*, *snort*, *snore*, …

  -ng is associated with noises:
  bang, *bong*, *clang*, *ding*, *ring*, *sing*, …

- In each case, a phonetic component (e.g. *gl-*, *sn-*) and a semantic component (e.g. ‘light’, ‘nose’)

Phonesthemes

- Origin of these patterns is obscure
- The words are not etymologically related
- The phonetic form is often sub-syllabic—not the sort of thing usually considered a morpheme in English (but see Rhodes and Lawler (1981)).
- Several analyses—morphemes, sound symbolism…
- Could they be merely coincidences in the lexicon? (Maybe there are enough *gl-* words in English that the ‘light; vision’ ones only a very small subset)

Definition of Phonestheme

- I adopt Bergen’s (2004) definition:

  (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. (293)

- Negative definition: not a phonestheme if we would otherwise predict the pairing (e.g. morphemes or etyma)
- Appeals to statistics: “better attested…than would be predicted”

- Phonesthemes
- ➢ **Psycholinguistic experiments**
- Statistical methods
- Procedure and results
- Closing Remarks

Psychological Reality

- Even without consensus about an analysis, experiments can still be performed
- Test psychological reality: do phonesthemes form a part of the mental grammars of speaker?
- If so, some effect on processing should be measurable
- Researchers have studied comprehension and production of phonesthemes

Hutchins (1998) and Bergen (2004)

- Hutchins: 46 English phonesthemes from a survey of the literature, asking participants to rate sound-meaning associations using questionnaires
- Bergen: morphological priming studies on *gl-* and *sn-*
- Both studies found effects: speakers do seem to have knowledge (conscious and unconscious) of the sound-meaning associations
- Clearly part of participants’ mental grammars

The Trouble with Experiments

- Phonesthemes are part of the mental grammar of speakers—but which phonesthemes?
- Chicken-and-egg problem: to evaluate phonesthemes, need phonesthemes to evaluate
- Experiments are expensive. It would be nice to have a method of finding candidate phonesthemes to test, or of validating the ones already proposed.
- In English, accumulated proposals at least give a starting point

- Phonesthemes
- Psycholinguistic experiments
- ➢ **Statistical methods**
- Procedure and results
- Closing Remarks

A Statistical Method

- Recall that Bergen’s (2004) definition was statistical
- Also did some simple counting in the Brown corpus:
  - 38.7% of word types and 59.8% of word tokens with *gl-* have meanings associated with light or vision
- Intuitively, a strong association. But what percentage is convincing rather than coincidence?
- A statistical method, based on concepts from Latent Semantic Analysis (LSA) (Deerwester et. al. 1990), document classification, and mutual information.

Term-document Matrix

- Consider a set of documents. Count the number of occurrences of each word and arrange in a matrix:

	the	*of*	…	*nose*	*light*	…
Doc 1	322	102	…	22	3	…
Doc 2	238	81	…	3	36	…
Doc 3	540	197	…	1	2	…
…						

- This matrix tells what words are associated with what documents

Document Classification

- Natural language processing technique
- Freely available BOW toolkit (McCallum 1996)
- Train a statistical classifier on two or more sets of documents (rows in the matrix)
- New documents are classified based on their similarity to documents in the training sets
- One way to gauge this similarity is **mutual information**

Mutual Information

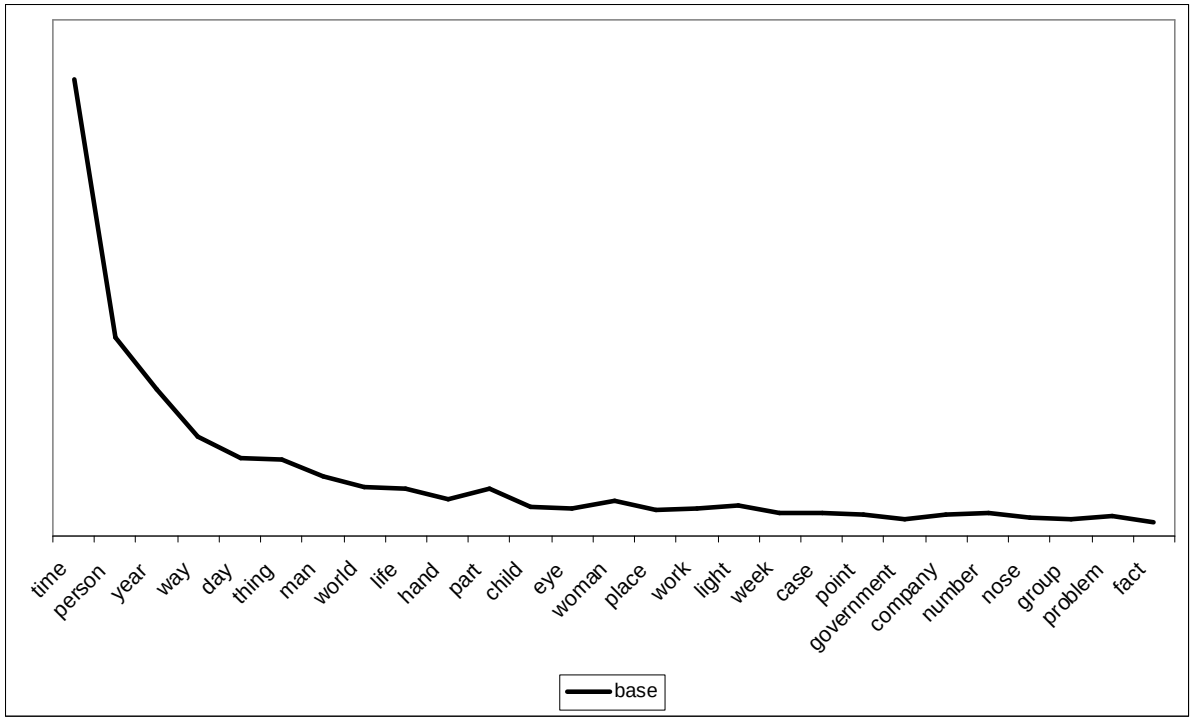
- From information theory. MI of two random variables is the amount of information knowing the value of one tells you about the value of the other.
- Formula: $$I(C;W_t) = \sum_{c \in C} \sum_{f_t \in \{0,1\}} P(c, f_t) \log \left(\frac{P(c, f_t)}{P(c)P(f_t)} \right)$$
- This can be calculated straightforwardly from the term-document matrix:
  - $P(c)$ = tokens in class c / total tokens
  - $P(f_t)$ = occurrences of some target word / total tokens
  - $P(c, f_t)$ = occurrences of target in class c / total tokens

Dataset

- To use them to examine phonesthemes, we need data we can view through the lens of these techniques
- A freely available English dictionary (1913 edition of Webster’s) processed to remove all formatting
- Treat each headword as a document whose content is its definition
- Look for form-meaning correlations: use orthography as a proxy for phonetic content, definition words as a proxy for meaning

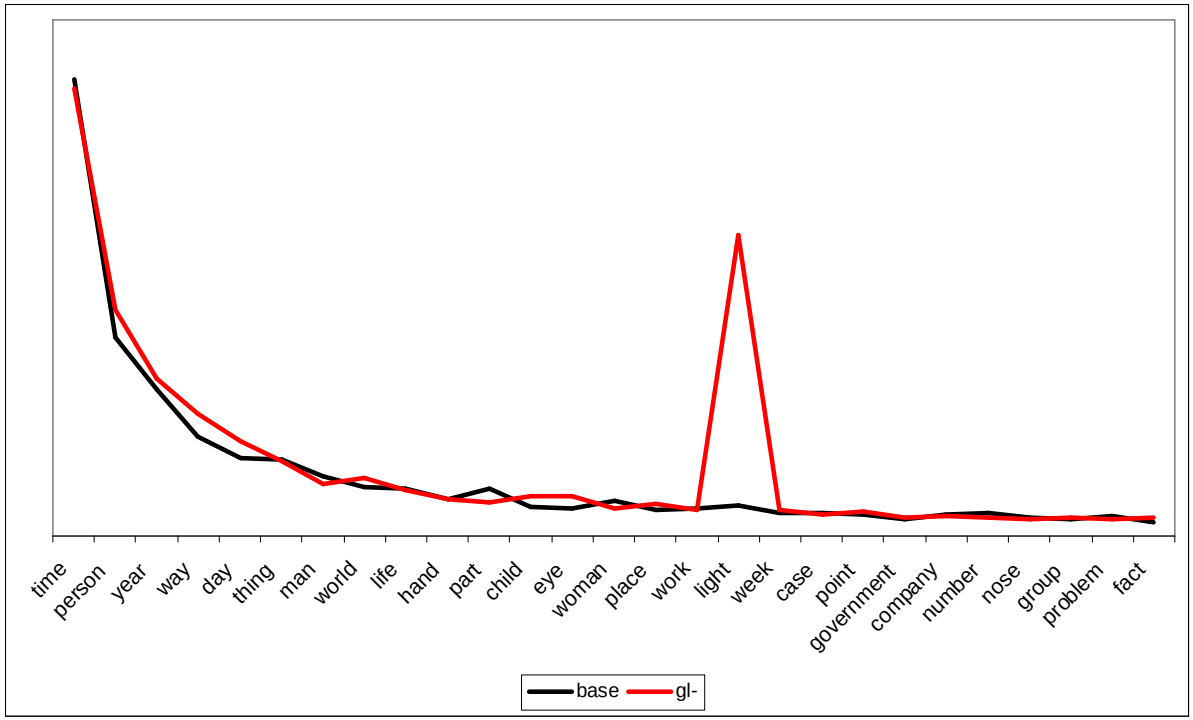
Form-meaning Pairings

- If phonesthetic meanings occur with greater than chance frequency, we should see this in the distribution of definition words:



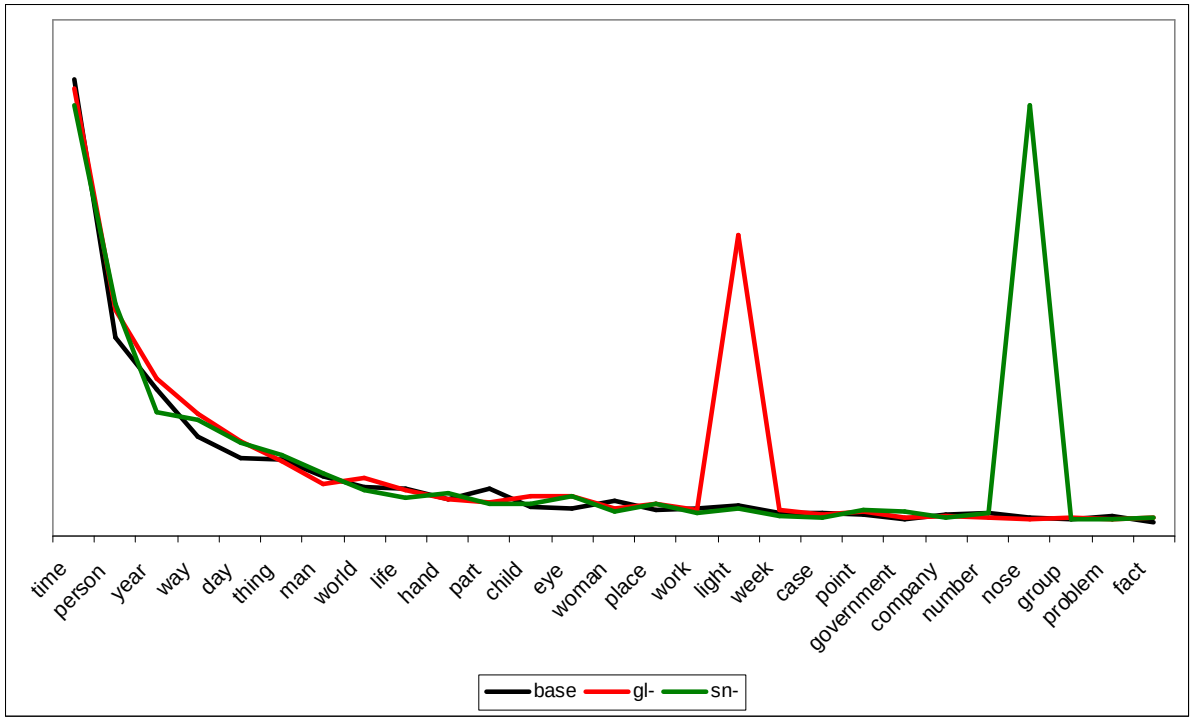
Form-meaning Pairings

- If phonesthetic meanings occur with greater than chance frequency, we should see this in the distribution of definition words:



Form-meaning Pairings

- If phonesthetic meanings occur with greater than chance frequency, we should see this in the distribution of definition words:



- Phonesthemes
- Psycholinguistic experiments
- Statistical methods
- ➢ **Procedure and results**
- Closing Remarks

Procedure

- Obtained and formatted a dictionary
- Treating definitions as documents, calculated the term-document matrix
- For each candidate phonestheme, considered two sets of definitions (rows in the matrix):
  - Headwords with the phonestheme’s phonetic form (e.g. all *sn-* words)
  - All headwords in the dictionary
- For each definition word, calculated the MI between two random variables:
  - Whether or not the word appears in a definition
  - Whether the definition belongs to the phonestheme class
- Sorted words by MI value and examine the most informative ones—if they have the phonesthetic meaning, that supports the candidate form-meaning correlation.

Sample Results

***sn-* ‘nose; snobbish’**

def. word	MI
nose	**0.0000565307**
sharp	0.0000163574
reprimand	0.0000133541
seize	0.0000121417
contempt	**0.0000119126**
short	0.0000118340
bite	0.0000116533
with	0.0000097613
laugh	0.0000097334
nasal	**0.0000090017**
angry	0.0000088951
check	0.0000087179
air	0.0000085600
nip	0.0000082975
catch	0.0000082894
fellow	0.0000082605
mucus	0.0000081098
surly	0.0000081098
rebuke	0.0000079575
mean	0.0000079168

***gl-* ‘light; vision’**

def. word	MI
smooth	0.0000232839
specious	0.0000222555
spherical	0.0000200744
look	**0.0000186537**
sullen	0.0000183769
light	**0.0000181011**
shine	**0.0000179517**
viscous	0.0000157358
bright	**0.0000121656**
luster	0.0000120111
ice	0.0000116167
stare	**0.0000114393**
acid	0.0000114003
comments	0.0000106663
sugar	0.0000101909
white	0.0000100298
and	0.0000088907
dilute	0.0000088024
vitreous	0.0000088024
commentator	0.0000086735

***st-* ‘firm; upright; linear’**

def. word	MI
to	0.0000340000
firm	**0.0000234677**
fixed	**0.0000201057**
in	0.0000138853
upright	**0.0000127493**
vessel	0.0000118034
walk	0.0000104120
precipitous	0.0000099669
post	0.0000094312
walking	0.0000093334
any	0.0000087957
antimony	0.0000086452
resolute	**0.0000085401**
position	0.0000081814
course	0.0000081642
spasmodic	0.0000079706
pointed	0.0000078469
obstinate	**0.0000077918**
cease	0.0000076854
thrust	0.0000076060

Two Tests for Significance

- Directly estimate confidence interval *p* (per word)
  - Apply the procedure once, then apply it 1000 more times with random sets of the same size as the candidate set
  - A word’s *p* = # times a random word had higher MI / 1000
- Estimate *p* based on rank of words (per phonestheme)
  - For *V* total word types and *w* words with the phonesthetic meaning, the chance of finding one or more in the top *n* is:

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

  - For $n = 20$, $p < 0.05$ if there are 68 or fewer words *w*

Significance

- The direct estimate can be calculated per word, but *p* is generally higher
- The rank test requires additional hypotheses (the words that will be considered “hits”), but produces much lower estimates of *p*
- Three kinds of phonesthemes, based on significance:
  - Strongly confirmed: both *p* values < 0.05
  - Weakly confirmed: only rank estimate < 0.05
  - Unconfirmed: neither estimate < 0.05

Results

- Of Hutchins’ 46 candidate phonesthemes:
  - 4 strongly confirmed: *sn-* ‘nose; snobbish’, *st-* ‘firm; upright; linear’, *-Vng* ‘ringing sound’, and *spr-* ‘to radiate out; elongated’
  - 33 weakly confirmed, including: *gl-* ‘light; vision’, *cl-* ‘noise from a collision’, *fl-* ‘motion, repeated or fluid’, str- ‘linear, forceful action’
  - 9 unconfirmed, including: *-am* ‘restrain in a small space’, *sm-* ‘insulting, pejorative term’, and *-ip* ‘quick movement or action’
- (See handout for the full results)

Strongly, Weakly, and Unconfirmed

***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

***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**

***-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

- Phonesthemes
- Psycholinguistic experiments
- Statistical methods
- Procedure and results
- **Closing Remarks**

Closing Remarks

- Described a technique for confirming the sound-meaning pairings underlying phonesthemes
- Given a dictionary and a hypothesis about the phonetic form of phonesthemes, possible to test all possible variants of that form
- Technique is language independent, but requires word segmentation and phonetic information. Some orthographies will be troublesome.
- Technique also finds morphemes and etyma. Unintended, but possibly useful.

References

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

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. 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.

McCallum, Andrew. 1996. BOW: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www-2.cs.cmu.edu/~mccallum/bow/

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.