Journal of Memory and Language 99 (2018) 122–150

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Weighing up the evidence for sound symbolism: Distributional properties predict cue strength

Chris Westbury[a,*], Geoff Hollis[a], David M. Sidhu[b], Penny M. Pexman[b]

[a] Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB T6G 2E9, Canada
[b] Department of Psychology, University of Calgary, Room AD255, 539 Campus Place N.W., Calgary, AB T2N 1N4, Canada

ARTICLE INFO

Article history:
Received 27 September 2016
Revision received 11 September 2017
Available online 20 October 2017

Keywords:
Sound symbolism
Semantics
Meaning
Psycholinguistics
Experimental psychology

ABSTRACT

It is well-established that there are relationships between word meaning and certain letters or phonemes, a phenomenon known as *sound symbolism*. Most sound symbolism studies have relied on a small stimulus set chosen to maximize the probability of finding an effect for a particular semantic category. Attempts to assign weights to sound symbolic cues have been limited by a methodology that has relied largely on forced contrast judgments, which do not allow systematic assignment of weights on the sound symbolic cues. We used a novel research approach designed to allow us to assign weights to sound symbolic cues. Participants made binary yes/no judgments about thousands of randomly-generated nonwords, deciding if they were good examples for each of 18 different semantic categories. Formal cues reliably predicted membership in several of those categories. We show that there is a strong inverse relationship between the average beta weight assigned to a phonological feature, phoneme, or letter, and the frequency of that cue. Our results also extend claims about the source of sound symbolic effects, by demonstrating that different poles of the same semantic dimension differ in their predictability from form cues; that some previously unsuspected dimensions show strong symbolic effects; and that features, phonemes, and letters may all contribute to sound symbolism.

Introduction

In his *Cratylus* dialogue, Plato gives Socrates the following words:

> "Must we not begin […] with letters; first separating the vowels, and then the consonants and mutes, into classes, according to the received distinctions of the learned; also the semivowels, which are neither vowels, nor yet mutes; and distinguishing into classes the vowels themselves? And when we have perfected the classification of things, we shall give them names, and see whether, as in the case of letters, there are any classes to which they may be all referred; and hence we shall see their natures, and see, too, whether they have in them classes as there are in the letters; and when we have well considered all this, we shall know how to apply them to what they resemble—whether one letter is used to denote one thing, or whether there is to be an admixture of several of them; just, as in painting, the painter who wants to depict anything sometimes uses purple only, or any other color, and sometimes mixes up several colors."
>
> [Plato, 360 BCE/1892]

A range of evidence (reviewed below) suggests that the sound symbolic associations Socrates was discussing, in which certain phonemes seem better suited as labels for certain meanings, are real, at least for some semantic categories. This evidence stands against one aspect of the "arbitrariness of the sign" (Saussure, 1916/1983), namely that no particular phoneme is more or less appropriate for any particular meaning. This is not a trivial matter, as arbitrariness is often taken as one of language's fundamental features (e.g., Hockett, 1963). Saussure (1916/1983) argued that there is no "reasonable basis" (p. 73) on which to discuss the appropriateness of a particular signifier for a signified, in contrast to our ability to discuss, for instance, "whether monogamy is better than polygamy" (p. 73). However, a variety of studies have shown that phonemes seem to have inherent associations with particular kinds of meanings, which suggests that such a discussion is viable and raises the possibility that these associations have affected language evolution (e.g., Berlin, 1994; Blasi, Wichmann, Hammarstrom, Stadler, & Christiansen, 2016; Johansson & Zlatev, 2013; Tanz, 1971; Ultan, 1978). Most notably, these

[*] Corresponding author.
E-mail address: chrisw@ualberta.ca (C. Westbury).

https://doi.org/10.1016/j.jml.2017.09.006
0749-596X/© 2017 Elsevier Inc. All rights reserved.

phoneme-meaning associations include the categories *sharp/round* and *large/small*, as we review in the next section. However, there is still uncertainty as to which particular phonological features, letters, or phonemes are the best predictors of these dimensions, how they are weighted, how they are related, and *why* they are the best, in part due to the methodological limitations of much previous work.

Several associated questions remain open: Is there any systematicity as to which linguistics features can act as sound symbols? How general are the categories in which sound symbolic effects can be found? Are the predictors of one pole of the dimensions the same (with reversed sign) as the predictors of the other pole, or are the poles separately symbolized? Are all predictors of semantic dimensions equal in their predictive force? Are the effects driven by phonological features, phonemes, graphemes, or some combination? Do biphones or bigrams contribute to the effects? If so, how strongly? Are all dimensions equivalent in their reliance on formal characters (phonological features, phonemes, or letters), or are different semantic dimensions differentially sensitive to these different formal characters? When evidence from features, phonemes, and letters are in conflict, is evidence from one stronger than from the other? Is the answer the same for all semantic dimensions? In this paper, we begin to address all these questions, using a large-scale study in which hundreds of participants classified thousands of nonword strings for their suitability as exemplars of 18 semantic categories.

Background

One of the best-known early demonstrations of sound symbolism was Köhler's (1929, 1947) report (building on closely-related observations made by Usnadze, 1924) that people had very strong intuitions about whether *maluma* (in 1947, or *baluma* in 1929)[1] or *takete* was a better name for a round or spiky shape. Köhler found that most of his participants thought that *maluma* was a better name for round things and *takete* a better name for spiky things, a result that has been much replicated (Davis, 1961; Holland & Wertheimer, 1964; Maurer, Pathman, & Mondloch, 2006; Nielsen & Rendall, 2011; Ramachandran & Hubbard, 2001; Sidhu & Pexman, 2016).[2] Sapir (1929), following up experimentally on an idea in Jespersen, 1925), undertook a related experiment in which vowels were manipulated. Sapir limited himself "to the meaning contrast 'large': 'small' as offering the most likely chance of arriving at relatively tangible results" (p. 226). Using a large number of participants, he showed that people preferred an open-front vowel *a* (/æ/) over a closed-front vowel *i* (/ɪ/) in labelling large things. This was later extended to the front-back dimension, with the finding that individuals preferred front vowels (e.g., /i/) when labelling small shapes, and back vowels (e.g., /ɔ/) when labelling large shapes (Newman, 1933).

The results of these studies have been frequently replicated. However, there are five problems with the reliance of the field on Köhler and Sapir's paradigm of forced choice experiments with a small number of contrasting strings pre-selected by the experimenter "as offering the most likely chance of arriving at relatively tangible results" (Sapir, 1929).

One problem is that such contrasts are often transparent manipulations, making it difficult to gather a lot of data per participant. Köhler could not reasonably have asked participants to make the same choice about the strings *kittatee* and *moolmer*, then *tikkikit* and *malinus*, and so on. Participants would have quickly caught on, and the task results would become uninformative. This is one reason why we have knowledge of the sound symbolic value of relatively few strings after nearly a hundred years of research into the phenomenon.

A second problem is that forced choice judgments are ambiguous in their interpretation, since they don't provide sufficient information for an outside observer to discern the basis of any choice. Concretely, when someone tells us that *mil* is a better name for a small thing than *mal*, we don't know if that person made the choice:

1. Because *mil* seemed like a good name for a small thing (a true positive) but *mal* had no interpretation either way (unclassifiable).
2. Because *mal* seemed like a good name for a large thing (a true positive) but *mil* had no interpretation (unclassifiable).
3. Because *mil* seemed liked a *poor* name for a large thing (a true negative) but *mal* seemed like a good name (a true positive).
4. Because *mal* seemed liked a *poor* name for a small thing (a true negative) but *mil* seemed like a good name (a true positive).
5. Because (as is usually assumed) *mil* and *mal* were both true positives in their respective categories.

These five possibilities are not exhaustive, because they do not take into account any considerations of *quantity*. Even in the fifth and normally assumed 'best case', we do not know from seeing a consistent response if *mil* and *mal* are *good* representatives of their respective categories, or (as we will argue from evidence later in this paper) mediocre representatives that differ just sufficiently to enable a consistent forced-choice to be made. As noted by Tukey (1969), it is very difficult to make theoretical scientific progress after merely noting *a difference* if we do not also quantify the *size* of that difference.

A third problem is that experiments relying on pre-selected contrasting strings are really experiments in 'intuition matching'. Köhler's and Sapir's experiments were not just a demonstration of sound symbolism; they were simultaneously a demonstration of the fact that other people share intuitions about sound symbolism. Of course, ultimately, sound symbolism must always rely on intuition, since the phenomenon is defined in terms of phenomenology. However, intuition matching in experimental design is scientifically dissatisfying for several reasons. One is that our conscious intuitions about a phenomenon may provide very little insight into the true nature of the phenomenon (consider, for example, the history of armchair theorizing about the nature of color prior to Newton). To the extent that a phenomenon is more complex or stranger than we are able to intuit, intuition matching experiments must leave us wondering how much of the phenomenon remains unexplained. By definition what remains to be explained lies outside of our intuitions. The second reason it is dissatisfying is that intuition matching also limits our ability to explain *why* the phenomenon of sound symbolism occurs. If we do not know why we *ourselves* feel that *k* is a good symbol of spiky things, asking other people if they feel the same way is not likely to shed any light on the question of why we all share that intuition. Much of the work on sound symbolism to date has served only to document that the phenomenon exists.

The last two problems with forced choice experiments are closely related to this intuition-matching problem. The fourth

[1] We have confirmed that the change from *baluma* to *maluma* occurred between the 2nd and 4th *printings* of the first American (1929) edition of Köhler's book. The 1st and 2nd printings occurred in the same month: April 1929. The 3rd printing (a copy of which we have not yet been able to locate) occurred in August 1929 and the 4th in October 1929. Since it is unusual to allow author edits of the same edition of a book between printings and we are aware of no comment by Köhler about this apparent change in his intuitions, we conclude that the early 1929 appearance of the nonword *baluma* was most likely *a printing error*. Thanks to Jan-Olaf Svantesson for making us aware that the change occurred between printings of the first edition rather than between editions (as implied in Westbury, 2005).

[2] We have omitted Westbury (2005) from this list as repeated failures to replicate the reported effect have cast that effect into doubt (see Westbury, in press).

problem is that these experiments necessarily focus our attention only on the most salient sound symbolism phenomena, i.e., on a small set of phonemes that have been intuitively recognized to give clear effects. They therefore offer no information about the range of sound symbolism effects that may exist, or about what constraints are in force on those effects. The final problem with forced choice contrasting strings experiments is a corollary of this: contrasting strings so strongly delimit the phenomenon of sound symbolism that it becomes impossible to generalize their results. We have intuitions only about a small number of the more extreme sound-meaning relationships and we study those few relationships again and again.

Evidence for all of these problems can be seen in reviewing the literature on experimental studies of sound symbolism to date. Ninety-nine sound symbolism experiments (drawn from 65 published papers) are summarized briefly in Table 1. Most sound symbolism experiments have relied upon a small number of stimuli. Sixty-one (62%) had eight or fewer stimuli or stimuli pairs. Twenty-seven (27%) followed Köhler in using just a single pair. Many experiments that used more than eight stimuli pairs nevertheless suffered from the limits of intuition matching by confining themselves to matching the famous intuitions of Köhler and Sapir. At least 40 (40%) of the experiments were focused on open/closed vowels, or closely related manipulations such as front/back vowels or low/high frequency sounds, like those first studied by Sapir and/or on studying *maluma*/*takete* (or nonwords deliberately chosen to be very similar, such as *uloomu*, *kipi*, and *moma*), following Köhler.

Theoretical considerations

Hinton, Nichols, and Ohala (1994) suggested that instances of sound symbolism fall into four main groups: *corporeal sound symbolism*, *imitative sound symbolism, conventional sound symbolism,* and *synaesthetic sound symbolism.*[3]

Corporeal sound symbolism refers to sounds such as grunts, moans, screams, and sighs, that attain their meaning (to slightly paraphrase Hinton et al., 1994, p. 2) more because they are *symptoms* than because they are *symbols*.

In *imitative sound symbolism* (onomatopoeia) words (such as *bowwow*, *meow*, and *cock-a-doodle-doo*) that are coined explicitly to mimic non-linguistic sounds in the world. Such sound symbolism is largely conventionalized (though see Rhodes, 1994, for a discussion of imitative sound symbolism 'in the wild') but that conventionality could play a role in linguistic sound symbolism by setting up conventions.

In *conventional sound symbolism*, associations between phonemes and meaning are assumed to be based on a clustering of particular word features with word meanings that is not attributable to any particular characteristics of the features, but rather due to a convention encoded in the lexicon. A well-known example is the existence of *phonesthemes,* morphologically-unrelated words with similar meanings and sounds (e.g., *glimmer, gleam, glint, glow, glare*), in which the shared *gl* has no apparent symbolic or morphological relationship to the semantic category of brief, intense visual experience. Baayen, Milin, Filipović-Đurđević, Hendrix, and Marelli (2011) proposed a causal mechanism that might allow for the spontaneous development of such pockets of form-meaning coherence. They note that "a discriminative learning approach [to language development] predicts that even small local consistencies in the fractionated chaos of local form-meaning correspondences will be reflected in the weights [for shared meaning between linguistic strings], and that they will codetermine lexical processing, however minute these contributions may be" (p. 56). They go on to remind us that Shillcock, Kirby, McDonald, and Brew (2001) found a reliable correlation between phonological and semantic distance of word pairs, proof that there are indeed 'small local consistencies in the fractionated chaos of local form-meaning correspondences' that could seed the growth of form-meaning systematicities (see also Blasi et al., 2016; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Perry, Perlman, & Lupyan, 2015). Conventional sound symbolism may thus be autopoietic (self-producing, in the sense of being independent of any other causes). This may be the case because some form-meaning relationships may arise purely by chance, which (following the assumptions of Baayen et al., 2011) allows that such systematic relationships are likely to arise, but has nothing to say about *which* relationships they will be. There may be no answer to the question *why does 'gl' associate with brief intense visual experiences?* except to say: because by chance it came to be that way. Of course, once phonesthemes exist, they may influence future sound symbolism.

A closely-related form of conventional sound symbolism may arise from the existence of semantic categories which share a small number of orthographic or phonological features for historical reasons. We consider one such example below, the category of flower names, which tend to share features due to the historical conventions of scientific taxonomy. We present evidence below suggesting, e.g., that 'heonia' is a better name for a flower than 'cruckwic'. We suspect most English speakers would agree even if they were not explicitly aware that 'ia' is widely used in English scientific taxonomy because it is used in Latin as a suffix for forming feminine nouns.[4] This theory of conventional sound symbolism is amenable to direct testing within some semantic categories, since it explicitly predicts that semantic categories should show sound symbolism effects to the extent that there are measureable phonological or orthographic regularities within words of that category. However, it is also limited to well-delineated categories: it is hard to imagine looking for conventional form regularities in the category of 'round things' or 'large things' since it would be very difficult or impossible to exhaustively list all the members of such broad and open-ended categories.

Hinton et al.'s (1994) final category of sound symbolism is *synesthetic sound symbolism*, in which some feature of the relevant phonemes leads to their association with particular kinds of stimuli, leading to a perceived link between the sounds and stimuli despite their being in separate modalities. As Von Humboldt (1836) put it, the sounds "produce for the ear an impression similar to that of the object upon the soul" (p. 73). This is the sort of sound symbolic association on which we focus here.

There are many proposals for how synesthetic sound symbolic associations might come to be. Following Sidhu and Pexman (2017), the causal mechanisms can be grouped into four broad categories (see also Spence, 2011, for a grouping of the mechanisms potentially underlying cross-modal correspondences, which Sidhu and Pexman note as an influence):

i. Association by a pattern of consistent co-occurrence between sounds and relevant stimuli in the environment.
ii. Association by shared properties between sounds and a stimulus.

[3] Note that Hinton et al. (1994) use the term sound symbolism to refer to relationships at the word level; thus far we have been using this term to refer to relationships at the phoneme level (i.e., between particular phonemes and certain meanings). These two uses are not in conflict; sound symbolic relationships between phonemes and meanings contribute to sound symbolic relationships at the word level, when a given word contains phonemes sound symbolically related with its meaning.

[4] Though we will present evidence below suggesting that there are other reasons why 'heonia' is a better name for a flower.

Table 1
99 Sound symbolism experiments.

Author[s] (Year)	Sound symbolic elements	Number of linguistic stimuli	Modality	Match Stimuli	Measure
Abelin (2015)	fl/kl/sp/kn/bl/sk/mj	8 NWs	Visual	Pictures of objects, NW phonaesthemes	Forced choice
	fl/kl/sp/kn/bl/sk/mj	8 NWs	Visual	Pictures of objects, NWs phonaesthemes	RT to decide if target word is a real Swedish word
Ahlner and Zlatev (2010)	Voiced/voiceless obstruents × front/back vowel	16 NW pairs	Auditory	Round/spiky shapes	Forced choice
Argo, Popa, and Smith (2010)	Reduplication	2 pairs of brand names	Visual/spoken aloud (or silent)/gustatory	Icecream	Rating scale
	Reduplication	3 pairs of brand names	Visual/auditory/gustatory	Icecream	Rating scale
	Reduplication	3 pairs of brand names	Visual/auditory/gustatory	Icecream	Rating scale
	Reduplication	[7 positively valenced pics., 7 negatively valenced pics.]	Visual	Pictures	Rating scale
	Reduplication	2 brand names	Visual	Hand sanitizer	Rating scale
	Reduplication	2 brand names	Visual	Phones	Rating scale
	Reduplication	2 brand names	Visual	Restaurant	Rating scale
Asano et al. (2015)	Kipi/moma	2 NWs	Visual and auditory	Round/spiky shapes	Infant visual preference
Athaide and Klink (2012)	Front/back vowel	4 pairs	Visual	Fictitious brand names	Forced choice
Auracher, Albers, Zhai, Gareeva, and Stavniychuk (2010)	Plosive/nasal	180 poems, 8 emotional expressions	Visual	Emotional expressions & poems	Rating scale
Aveyard (2012)	Front/back vowel	32	Auditory	NWs	Forced choice
Baxter, Ilicic, Kulczynski, and Lowrey (2015)	Front/back vowel	2 brand names	Visual	Image of toy, brand name (Illy/Ully) inside a triangle or ellipse logo	Rating scale
Baxter and Lowrey (2011)	Front/back vowel	4 word pairs	Visual	Ice cream brand names	Forced choice
	Front/back vowel	4 toy names	Visual	Toys, toy names	Forced choice
Bentley and Varon (1933)	CVC NWs	10	Auditory	Large/small	Free association
	CVC NWs	10	Auditory	10 categories	Forced choice
	CVC NWs	5	Auditory	5 categories and their opposites	Forced choice
	CVC NWs	36 pairs	Auditory	3 categories and their opposites	Forced choice
Bremner et al. (2013)	Bouba/kiki; water; chocolate	2	Visual, gustatory, auditory	Round/spiky shapes	Forced choice
Brown, Black, and Horowitz (1955)	Foreign antonym pairs	21 pairs	Auditory	Foreign words	Foreign word to English word matching
Coulter and Coulter (2010)	Front/back vowel & fricatives/stops	8	Visual	Product prices	Rating scale
Cuskley (2013)	CVCV NWs	105	Auditory	Motion of a circle	Speed adjustment
Cuskley et al. (2015)	CVCV NWs	8	Visual	Round/spiky shapes	Rating scale
	CVCV NWs	8	Auditory	Round/spiky shapes	Rating scale
Davis (1961)	Oloomu/takete	2	Auditory & visual	Round/spiky shapes	Forced choice
D'Onofrio (2014)	Labial/Alveolar/Velar x Voiced/Voiceless	12	Auditory	Round/spiky shapes	Forced choice
	Front/back vowel	12	Auditory	Round/spiky kitchen items	Forced choice
Doyle and Bottomley (2011)	Front/back vowel	2	Visual	Frish/Frosh	Rating scale
Favalli, Skov, Spence, and Byrne (2013)	Overall gestalt	3 pairs of NWs, 11 abstract associations, 4 category names	Gustatory and visual	Danish open-faced sandwiches, NWs, abstract associations, category names	Ranking
Fenko, Lotterman, and Galetzka (2016)	Ramune/Asahi	8	Visual	Visual	Rating scales
Flumini, Ranzini, and Borghi (2014)	k/t vs g/m	24 object, 8 NWs	Visual	Black & white line figures	Forced choice
Fort, Weiß, Martin, and Peperkamp (2013)	CVCV NWs	28	Auditory	Round/spiky shapes	Infant visual preference
	CVCV NWs	12	Auditory	Round/spiky shapes	Infant visual preference
	CVCV NWs	12	Auditory	Round/spiky shapes	Infant visual preference
Holland and Wertheimer (1964)	Baluma/takete	2	Visual	Round/spiky shapes	10 rating scales
	Maluma/takete/kelu	3	Visual	Round/spiky shapes	Rating scale

(*continued on next page*)

Table 1 (*continued*)

Author[s] (Year)	Sound symbolic elements	Number of linguistic stimuli	Modality	Match Stimuli	Measure
Imai, Kita, Nagumo, and Okada (2008)	[fast/slow, heavy/light]	6 mimetics, 12 videos	Auditory for mimetic, visual for video, live speaker for mimetic in Experiment 2	Characters walking	Rating scale
Imai et al. (2015)	Kipi/moma	2 NWs, 2 shapes	Visual and audio	Round/spiky shapes	Infant visual preference
Irwin and Newland (1940)	NWs	10	Visual	Complex shapes	Forced choice
Kantartzis, Imai, and Kita (2011)	NWs	4 altered Japanese mimetics, 4 with structure of English verbs	Visual	8 NWs describing movement	Indicate preference
Kawahara and Shinohara	Voiced/voiceless obstruents	40 test, 20 filler	Auditory	NW	Rating scale
Klink (2000)	i e/o u, f v/g k, p t/b d, f s/v z	124	Visual	Product names	Forced choice
Klink (2001)	NWs	3 brand names * 3 products	Visual	Brand names	Rating scale
Klink (2003)	i/o, i/u, e/o, e/u, f/v, g/k	4 brand name pairs	Visual	Brand names	Forced choice
	Vowel/consonant	2 brand names * 2 brand mark	Visual	Brand names & brand marks of beer	Rating scale
Klink and Athaide (2012)	Front/back vowel	4	Visual	Fictitious brand names	Rating scale
Klink and Wu (2014)	NWs	20 names in five groups	Visual	Brand names	Rating scale
	High/low frequency vowel/consonant	16 names in 4 groups	Visual	Brand names	Rating scale
	High/low frequency vowel/consonant	16 names in 4 groups	visual/auditory/gustatory	brand names	Indicate preference between pairs for speed and size
Kovic and Plunkett (2009)	Mot/ riff, dom/shick	9 training, 140 test	Visual and auditory	Animal-like shapes	Forced choice RT
Kuehnl and Mantau (2013)	Front/back vowel	4 pairs	Visual	Bran name preference	Forced choice
LaPolla (1994)	English antonym pairs	40	Auditory	English antonym pairs	Forced choice
Lowrey and Shrum (2007)	Front/back vowel	10 word pairs	Visual and auditory	NW brand names	Forced choice
Lupyan and Casasanto (2015)	Foove/crelch	2 names, 24 aliens	Visual	Two visual “alien” species	Forced choice
Maglio, Rabaglia, Feder, Krehm, and Trope (2014)	Front/back vowel	6	Visual	Fictitious city names, aerial view of rural landscape	Divide landscape into regions
	Front/back vowel	2 words	Visual	Actions	Forced choice
	Front/back vowel	2 words	Visual	Situational descriptions	Rating scale
	Front/back vowel	2 words	Visual	Situational descriptions	Rating scale
	Front/back vowel	2 words	Visual	Situational descriptions	Rating scale
Maurer et al. (2006)	Front/back vowel × k/b	4 NWs	Visual and auditory		Forced choice
Miyazaki et al. (2013)	Kipi/moma	2 shapes, 2 names	Visual and auditory	NW names for spiky & round shapes	Infant visual preferenece
Monaghan et al. (2012)	Plosive vs. non-plosive consonants	16 NWs, 16 shapes	Auditory	NWs & shapes	Forced choice
Myers-Schulz, Pujara, Wolf, and Koenigs (2013)	Positive/negative affect of pictures	Visual test: 35 NW pairs; Auditory test: 22 NW pairs.	Visual and auditory	NW phoneme strings & pictures	Forced choice
Ngo and Velasco (2012)	Akete-maluma, bouba-kiki	6	Gustatory	Exotic fruit juices	Rating scales + Color decision
Nielsen and Rendall (2011)	Consonant/vowel placement	30 names	Visual	Round/spiky shapes	Forced choice
	Consonant/vowel placement	42 word pairs	Visual and auditory	Round/spiky shapes	Forced choice
Nygaard, Cook, and Namy (2009)	Antonyms	21 antonym pairs	Auditory	Japanese antonym pairs	Forced choice
O’Boyle, Miller, and Rahmani (1987)	Uloomu/takete	2	Auditory	Shapes	NW-shape matching
Ohtake and Haryu (2013)	Front/back vowel	2 pairs of brand names	Visual and auditory	Disks	Size judgment
	Front/back vowel (mouth shape)	2 NWs, 2 shapes	Oral and visual	grey disks, mouth shape	Size judgment
Ozturk, Krehm, and Vouloumanos (2013)	Front/back vowel & k/b	2 NWs, 2 shapes	Visual and auditory	Curvy/angular	Infant visual preferenece
Parault and Schwanenflugel (2006)	Obsolete English words	40 words	Visual	Sound-symbolic & non-sound symbolic obsolete English words	Word knowledge
Parise and Spence (2012)	mil/mal	2	Auditory	Large/small circles	Forced choice
	Takete/maluma	2	auditory	Round/spiky shapes	Forced choice
	High/low sine wave	2 sounds	Auditory	Large/small circles	Forced choice

	High/low sine wave	2 sounds	Auditory	Angled lines	Forced choice
	High/low square wave	2 sounds	Auditory	Straight/curvy lines	Forced choice
Parise and Pavani (2011)	Spectral analysis of vowel production	1 vocal production	Vocal production	Regular convex polygons	Loudness of utterances
Park and Osera (2008)	F1/F2 frequencies	3 brand name pairs	Visual	Categories of products & artificial brand names	Forced choice
Peña, Mehler, and Nespor (2011)	i/o	2	Auditory	Large/small circles	Infant visual preference
	e/a	2	auditory	Large/small circles	Infant visual preference
Reilly et al. (2012)	English word characteristics	100 NWs	Auditory	NWs	Forced choice
	Syllable length, vowel duration	20 NWs	Auditory	NWs	Forced choice
	Morphological complexity	80 words	Visual	English nouns	Forced choice
Roblee and Washburn (1912)	Nonword VC strings	~208 [less real English VC words]	Auditory	VC	Rating scale
Rogers and Ross (1975)	Takete/maluma	2	Auditory	Round/spiky shapes	Forced choice
Sapir (1928)	Front/back vowel	60	Auditory	Word & NW pairs	Forced choice
Shinohara and Kawahara (2010)	Voiced/voiceless vowels	40	Visual	Nonword VC strings	Rating scale
Shrum and Lowrey (2012)	Front/back vowel	6 pairs	Visual	NW pairs	Forced choice
Sidhu and Pexman (2015)	Stops/continuants & front/back vowels	5 round-sounding male names, 5 round-sounding female names, 5 sharp-sounding female names, 5 sharp-sounding male names, 20 pairs of alien-like character silhouettes which are round or sharp	Visual	Proper names & alien shapes	Name choice
	Stops/continuants & front/back vowels	10 pairs of round/sharp male names, 10 pairs of round/sharp female names	Visual	Proper names & adjectives	Name choice
Simner, Cuskley, and Kirby (2010)	Tastes x concentration	[Continuous sound sliders]	Auditory	Sound qualities	User-selected sound qualities
Walker et al. (2010)	Sliding whistle	[Sliding whistle]	Visual	Ball motion	Infant visual preference
		[Sliding whistle]	Visual	Morphing shape	Infant visual preference
Westbury (2005)	Stop/continuant consonants	60 NWs and 60 words	Visual	Spiky/curvy	Forced choice RT
Yorkston and Menon (2004)	Front/back vowel	2	Visual	Product adjectives	Rating scale

iii. Association by overlap in the ways in which sounds and stimuli are neurally coded.
iv. Association by evolution.

In the case of co-occurrence, associations may be due to a particular linguistic feature (e.g., high pitch) and instances of a particular meaning co-occurring in the environment. For example, one might explain the association between closed vowels (which in general are perceived to have higher pitch; Ohala & Eukel, 1987) and small sizes with reference to the fact that small entities tend to emit higher pitches (Fitch, 1997; Ohala, 1983, 1984). Ohala (1994) theorized that sensitivity to this co-occurrence might have become innate through evolutionary processes–an example of an evolved sensitivity to sound symbolism (i.e., mechanism four).

In the case of sound symbolism due to shared properties, the features of phonemes are assumed to have some property (i.e., perceptual, conceptual, or affective) in common with associated stimuli. For instance, Bozzi and Flores D'Arcais (1967) found that round- and sharp-sounding nonwords shared certain connotations with the round and sharp shapes with which they are associated.

Finally, some have posited that sound symbolic associations might arise from the way information is coded in the brain. Ramachandran and Hubbard (2001) suggested that there may be a close analogy between synaesthesia and sound symbolism, and presented some evidence that the former may be due to neural cross-talk. They speculated that sound-shape symbolism might be due to interactions between visual representations in the inferior temporal lobe, and sound representations in primary auditory cortex. Westbury (2005) speculated that they might occur in the left mid-fusiform gyrus, which is known to be involved in both form and word processing, or the left lateral posterior temporal lobe, which has been implicated in synaesthesia involving words (Paulesu et al., 1995). However, there is no hard evidence to support any of these speculations. Vainio, Schulman, Tiippana, and Vainio (2013) proposed a related idea, suggesting that the link between certain vowels and sizes might arise from links between articulation and grasp neurons (see also Kovic, Plunkett, and Westermann, 2010).

Of course, these theories need not be mutually exclusive. Multiple mechanisms may contribute to a single instance of sound symbolism; it is also possible that different instances of sound symbolism are explained by different mechanisms. In addition, a question that has not been addressed is whether certain *kinds* of linguistic elements might be more likely than others to have sound symbolic associations.

There is clearly much work to be done exploring the mechanism that underlies sound symbolism. One challenge is that researchers have often been forced to derive theories based on somewhat restricted datasets. This is because much of the existing literature consists of studies focusing on specific subsets of language, and their associations with a small number of semantic dimensions. In addition, many theories have been generated to explain a *specific instance* of sound symbolism. In the present study, we used a novel approach that avoids these issues. Instead of only examining particular phonemes and semantic dimensions, we surveyed all of English phonology for associations with a variety of semantic dimensions. This was used to explore a potential over-arching factor in sound symbolism that has not been previously examined: the distributional properties of sound symbolism cues.

A necessary goal for developing predictive theories is not just to identify the phonological features, phonemes, and graphemes that play a role in sound symbolism without relying on contrastive judgments which mask their mechanism, but to try to *weight* those elements in order to identify the degree to which sound symbolization for a particular category depends on features, phonemes, or letters, or some combination. Our motivation for this approach comes from a recent theory of sound symbolism that was used to make and test explicit hypotheses. Westbury, Shaoul, Moroschan, and Ramscar (2016) studied sound symbolism related to the perceived humor in nonwords. Following earlier theories of humor, they suggested and showed that this form of sound symbolism was predictable as a function of *the statistical properties of the strings*, rather than the explicit content (i.e., the specific letters and phonemes) of the strings. They were able to predict funniness ratings and forced choice decisions for nonwords, with good accuracy across the entire range of judged funniness. They did so by computing a simple measure of string probability related to Shannon entropy (Shannon, 1948), computed across the probabilities of the letters in the strings without regard to what letters contributed to the probabilities in any particular string.

The theory that there are statistical/information theoretic factors contributing to sound symbolism has been little explored, but makes good *a priori* sense. Sound symbolism is essentially a signal detection problem, in which a person needs to classify strings into two categories (sound symbolic or not). That classification depends on the detection of a (possibly complex) signal that we know to be rare, since (as we demonstrate below) most nonwords do not have any sound symbolic properties. Strings that are most clearly perceived as having sound symbolic properties must carry the most clearly discriminative cues to the presence of that detected signal. This allows us to make some specific predictions about the relevant cues. One is that rare sounds or letters are more likely to make good discriminative cues than common sounds or letters, precisely because they are rare. Common cues, in virtue of being common, cannot also be discriminative, since they will by definition appear in (i.e., not serve a cue to discriminate between) many strings. For example, consider trying to identify a car from knowing that it has a '1' in its license plate (a common cue) versus trying it identify it by knowing its actual license plate number (an uncommon cue). The common cue is (*in virtue of being common*) a weakly discriminating signal, whereas the uncommon cue is (in this case) a perfectly discriminating signal, since it picks out exactly one exemplar from the class of cars. A corollary of this inverse relationship between frequency and discriminatory utility is that phonological features, most of which are ubiquitous by their very nature, can be expected to be weak cues to sound symbolism, because they occur too often to be useful as a discriminatory cue. We can also predict that cues that are unambiguous (i.e., have a consistent orthography-phonology mapping) will better able to serve as cues than ambiguous cues, since ambiguous cues are by definition poor cues. This has several corollaries. One is that it is a second strike against vowels as cues to sound symbolism (the first strike against them being that all vowels are common), especially when those vowels are represented as letters rather than phonemes, because vowels-as-letters tend to have more than one pronunciation and are thus more likely to have ambiguous value as signals. It is also predictable that, *pari passu*, a cue with a single interpretation (i.e., a letter that maps onto only one phoneme) should be a better cue than a similar cue that has more than one interpretation (i.e., a letter that maps onto more than one phoneme).

Note that, except insofar as predicting that good cues will have certain properties (i.e., be rare and unambiguous in proportion to their utility as cues), none of these predictions is about any *specific* letter or phoneme. However, the predictions do suggest that it is likely that only a small set of cues (rare and unambiguous letters or rare phonemes) will be suitable as strong cues to sound symbolism. This is turn limits the number of sound symbolic categories that are possible in a language. For example, a cue cannot be a signal of both large and small size, since for a cue to be both a signal and a non-signal is useless. We can therefore expect that the small number of cues that can serve should be distributed with little overlap between different sound symbolic categories, and that overlap should only occur when the categories are not contradictory (i.e., perhaps between the categories of *small* and *feminine*, as hypothesized by Jespersen (1925), or between the categories of *large* and *round*, since largeness in animals and people is often accompanied by roundness for biological reasons).

These are all testable predictions, but in order to test them we need to be able to assign weights to sound symbolism cues in different semantic categories. To this end, in the present study, we made several design choices that were different than those used in much of the existing literature on sound symbolism. In contrast to many of the previous studies, we did not start by defining our stimuli using either our own or anyone else's intuitions about sound symbolism. Instead we constructed a large number of nonword strings randomly, using almost all possible English phonemes. By building regression models over the cues contained in each string, we are able to assign weights to the 27 features, 41 phonemes, and 26 letters that might signal a string's membership or non-membership in a given category. We asked participants to make judgments about whether nonwords were good names for members of 18 different categories, described in more detail in the methods section below. In addition to testing the predictions we derived from the signal detection conception of sound symbolism, we used this large-scale data set to explore the various questions related to sound symbolism that were mentioned at the outset.

Method

Participants

Participants were 214 students (51 [23.7%] males; 210 [98.6%] self-reported right-handers). All claimed to be native English speakers, defined as having learned to speak English before the age of 5. They had an average [SD] age of 19.9 [2.84] years and an average [SD] of 13.9 [3.8] years of education. They participated in return for partial course credit at either the University of Calgary [71 participants] or the University of Alberta [143 participants]. Forty-four participants at the University of Alberta participated in this experiment as well as in two other unrelated very simple experiments, with this experiment always given last. The remainder participated only in this experiment.

Stimuli

We selected 21,220 English words that were not morphologically inflected (though not necessarily monomorphemic) and that were marked, using ASCII characters and capitalization, for syllabification, stress, and word onset. For example, this list included the entries #AL-go-RITHM#, #KI-OSK#, #a-BODE# and #ZOOM#.[5] We used the software program LINGUA (Westbury, Hollis, & Shaoul, 2007) to create 629,767 nonwords from this list of words, using a Markov chaining process that chained on three characters, eliminating real words in its English dictionary. This process guarantees that any three contiguous characters in a nonword also appeared in a word and samples the trigrams proportionally to their probability of occurrence in real words. This Markov chain process created nonwords that were syllabified, marked for one level of stress, and that were guaranteed to have real word beginnings. Word endings are not guaranteed by Markov chaining, since the algorithm merely selects a statistically-weighted allowable continuation for any three characters.

This method produces nonword strings that are highly plausible as English words and that have highly plausible syllable boundaries. For example, the output list generated by LINGUA contained nonwords such as #DOL-e-FOR, #CAL-clar#, #FIR-ric#, and #in-TU-za#. These strings illustrate the necessity of generating syllable boundaries within multi-syllabic nonwords, and of presenting auditory strings in experiments focusing on the form effects of nonwords. The orthographic strings *dolefor* and *intuza* have ambiguous pronunciations without syllabification information, like many English words. For example, the English word *judaica* might plausibly be read with three syllables as /dʒuː-dɛɪ-kə/ [JOO-DAY-ka] if one did not know it has four syllables. From their orthographic representation alone *dolefor* and *intuza* might plausibly be pronounced /dəʊl-fɒr/ [DOLE-FOR] (if the syllabification is interpreted as #DOLE-FOR#) or as /ɪn-tʌz-ə/ [in-TUZZ-a] (if syllabified as #in-TUZ-a#).

From the large set of generated nonwords we hand-selected 12,556 nonwords that ranged from three to eight letters in length and that were deemed to have an unambiguous pronunciation, given their marked syllabification and stress pattern, and to have plausible word endings. We then converted these nonwords to their phonological representation using custom-written rule-based software that converts syllabified orthographic strings such as ours into a written English phonological code (Derwing, Priestley, & Westbury, in preparation; for methodological details from closely-related work see Derwing & Priestly, 1980; Derwing, Priestly, & Rochet, 1987).

We used Apple's text-to-speech software to convert each of these phonological representations into an .AIFF sound file (see Table 2 for an overview of Apple's phonological system and its relationship to IPA). Although usually correct, these representations were taken as starting points, not ending points: at least one author (most often two) listened to each sound file to make sure that the file was not erroneous, distorted, or otherwise problematic. Strings with problematic files were fixed by hand, being either phonologically improved and regenerated, or discarded, until we had 8000 nonword strings with sound files that were deemed by at least one author to be correctly and clearly enunciated. These files will be made available at: http://www.psych.ualberta.ca/~westburylab/publications.html.

After all data had been collated, we noticed that we had included no nonwords containing a voiced *th* (/ð/). This phoneme occurs in only a few words in English: mostly in high frequency words that are irregular because of this rare voicing (e.g., *that, than, the, they*) but also systematically in words ending in *-ther* (e.g., *feather, mother, gather, leather*), systematically in words that end with *-the* and their derivations (e.g., *breathe/breather, bathe/bather, seethe/seething*), and uniquely in the English word *rhythm*. Four of the 8000 nonwords (*cothe, flathe, snathe*, and *cythm*) fit a pattern for voiced *th* /ð/ but had been transcribed and recorded with an unvoiced *th* /θ/. Together these nonwords had been used 22 times in the experiment, constituting 0.05% of all stimulus presentations in the experiment. Because they met a pattern for voicing but had not been voiced when presented to participants, we removed these four nonwords from further consideration, leaving us with 7996 unique strings for analysis. The experiment was therefore analyzed using 7996 strings and sound files (average [SD] orthographic length: 6.1 [1.0], range: 3–8; average [SD] number of syllables: 1.9 [0.6], range: 1–4).

As outlined in detail in the following section, we asked participants to make a yes/no decision about whether a string was a good exemplar for each of 18 categories.[6] The 18 category labels we used

[5] The original representation additionally also marked primary stress in words with more than one stressed syllable, but since the Markov chain process has no memory (i.e. cannot know if it has already marked the primary stress in a generated string), we did not use this information.

[6] Participants actually made decisions about 20 categories. In response to reviewer feedback (and *pace* Lee Wurm) in this paper we report results for only 18 of those categories, disregarding the terms that anchor the *ad hoc* dimension *dangerous/safe*.

Table 2
Correspondence of Apple's phonological representation to the International Phonetic Alphabet. 41 phonemes were used in this study. Consonants not listed here are represented in both systems by their standard orthographic forms.

Example	Apple	IPA	Stress
c**a**t	AE	æ	
m**a**te	EY	ɛɪ	
r**o**ck	AA	ɒ	
c**au**ght	AO	ɔː	
s**ee**d	IY	iː	
s**i**de	AY	aɪ	
s**i**ck	IH	ɪ	Stressed
ros**e**s	IX	ɪ	Unstressed
b**e**d	EH	ɛ	
c**oa**t	OW	əʊ	
f**oo**d	UW	uː	
c**ow**	AW	aʊ	
c**oy**	OY	ɔɪ	
c**u**t	UX	ʌ	Stressed
about	AX	ə	Unstressed
f**oo**t	UH	ʊ	
h**ur**t	AXr	ɜː	
thin	T	θ	
then	D	ð	
shoe	S	ʃ	
a**z**ure	Z	ʒ	
cheese	C	tʃ	
jump	J	dʒ	
si**ng**	N	ŋ	
yes	y	dʒ	

are shown in Fig. 1, which we discuss in more detail below. Six of those labels were chosen because they are anchors of a polar dimension that has previously been associated with sound symbolism (*large/small, sharp/round, feminine/masculine*). The remaining 12 were nouns chosen because previous studies had found them to be rated at the extreme ends of valence (Warriner, Kuperman, & Brysbaert, 2013) and concreteness (Brysbaert, Warriner, & Kuperman, 2014), two features which have been argued to play a central role in the structure of semantics (Hollis & Westbury, 2016). In particular, we chose three nouns for each of the four categories: high valence/high concreteness (*flower, gem, toy*), high valence/low concreteness (*wisdom, spirituality, virtue*), low valence/high concreteness (*wasp, bomb, fungus*) and low valence/low concreteness (*sadness, fraud, injustice*). In the Warriner et al. study from which the valence ratings were drawn, valence was rated on a nine-point scale; we conducted an independent samples *t*-test to confirm that low valence nouns ($M = 2.48$, $SD = 0.26$, range: 2.05–2.79) had a significantly lower rating than high valence nouns ($M = 7.31$, $SD = 0.39$, range: 6.70–7.94), $t(10) = 25.05$, $p < .001$. In the Brysbaert et al. study from which the concreteness ratings were drawn, concreteness was rated on a five-point scale; we conducted an independent samples *t*-test to confirm that low concreteness nouns ($M = 1.52$, $SD = 0.26$, range: 1.07–1.82) had a significantly lower rating than high concreteness nouns ($M = 4.87$, $SD = 0.15$, range: 4.59–5.00), $t(10) = 27.31$, $p < .001$. For analyses, nonword judgments for nouns in each category were combined to allow us to examine the polar dimensions of *valence* and *concreteness* (e.g., examining the phonemes that were associated with high valence items).

Procedure

Participants were seated in front of a computer. After answering some simple demographic questions, they were given the following written instructions:

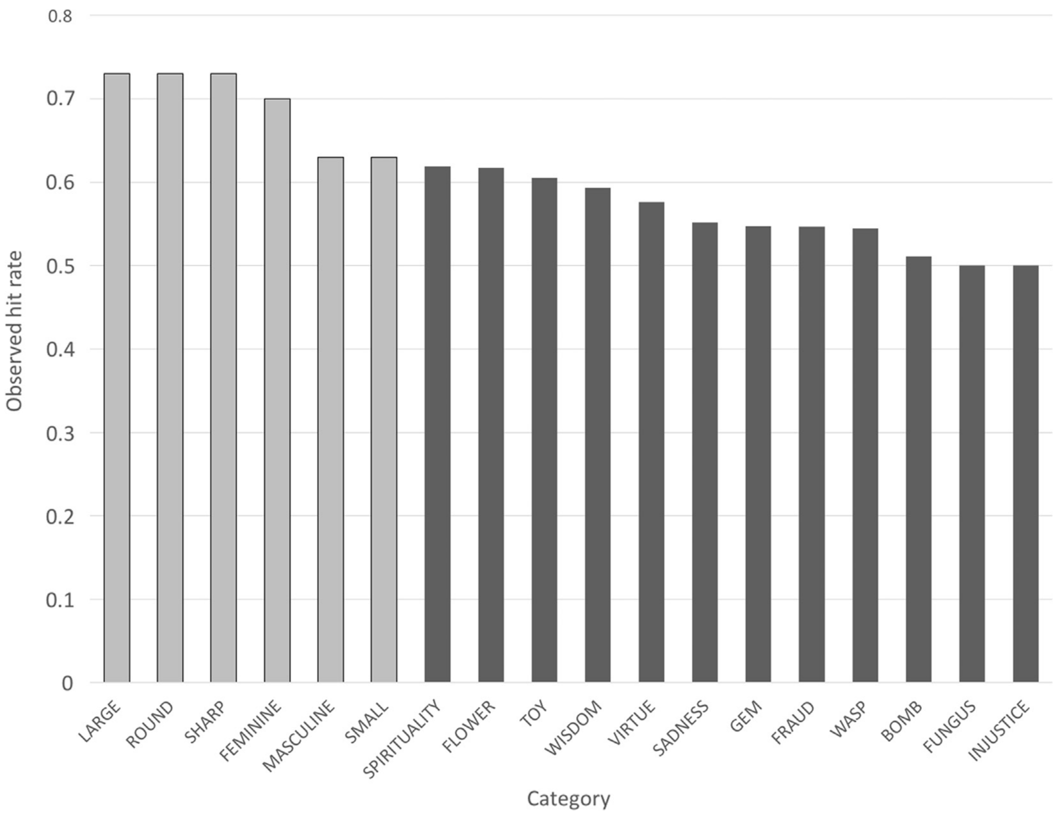

Fig. 1. Hit rates (agreement with human decisions) for 18 semantic categories. The six categories in grey reflect 'high confidence' judgments on the subset of decisions which had estimated membership probability in one pole (e.g., *large*)- estimated membership probability in its opposing pole (e.g., *small*) > 0.30. Hit rates for all except the last three categories are reliable with $p < .05$. See the associated tables for more details on how these hit rates were achieved (i.e., true positive and true negative rates).

> "We are interested in how people decide what makes a good word in English. In this experiment, we are going to ask you to make simple yes/no decisions about whether a nonword string might make a good word in English for a particular category.
>
> We will show you a category name on the screen. For example, we might show you 'A name for a vegetable'. Shortly after the category appears you will see a nonword printed in larger letters above that category name. At the same time, the computer will pronounce the nonword. If you think it would make a good word for that category, hit the 'c' key, for 'correct'. If you think it would not make a good word or are not sure or have no opinion, hit the 'x' key, for 'incorrect'.
>
> There are no right or wrong answers; we are just interested in your gut instinct. There is no requirement that you accept or reject equal numbers of words so do not let your decision be influenced by your previous decisions. Just decide about each string by itself.
>
> We will start with two examples to get you used to the task."

Participants were then asked to put on headphones through which the stimuli were presented aurally. The two examples that followed were nonword strings and category labels that were not used in the experiment, asking (in randomized order) whether the nonword *dalmul* was a good name for *A type of car* and whether the nonword *telch* was a good name for *A vegetable*. After the two examples were shown, participants were asked if they had any questions. When they were satisfied that they had none, they were asked to begin the experiment, and were left alone to complete the task.

The stimuli were presented using custom-written software. The category names were presented in 36-point dark grey ('Gray30'; RGB 77,77,77) Times font for 750 ms. After that, participants saw a '+' above the category name to orient them, replaced 250 ms later by the presentation above the category name of the nonword in 64 point black Times font which was presented simultaneously in the auditory modality. The category name stayed on screen when the nonword was presented, and both the category name and the nonword stayed on screen until the participant made a legal key press, in order to eliminate any reliance on memory. The ITI was 1000 ms. After 100 trials, the participants were given a self-timed break. We inserted this break (and cut the number of stimuli presented from our originally-intended 400 to 200) because pilot testing suggested that the task became difficult and frustrating after participants had made many decisions.

Every participant was required to make ten decisions in all 20 of the original categories, or 200 decisions in all. Since we had 214 subjects, we obtained 2140 decisions per category. It is key to this experiment that we started with 40 times more nonword stimuli (8000) than any individual participant saw (200), and that each participant's file was randomly generated for that participant (although the original 8000 nonwords were selected without replacement until they were all used, so each of the 8000 nonwords was used equally often). These experimental features are key because it is by their means that we separated particular nonwords from particular semantic categories. No single nonword was ever presented more than six times across the entire duration of the experiment (each of 215 participants saw 200 [1/40th] of 8000 stimuli, so we ran through the stimulus set $215/40 = 5.375$ times; i.e., $0.375 * 8000 = 3000$ strings were used six times). Almost all (91.7%) of the 7996 nonwords were judged just a single time, across all participants, within any semantic category. Eight percent were judged twice, and just 0.3% were judged as many as three times. Given this near-complete separation of nonwords from categories, any findings of a consistent relationship between nonword form and semantics must be due to the features, phonemes, or letters in the strings, and not to any other characteristics such as, for instance, resemblance of nonwords to particular exemplars of or labels for a semantic category (i.e., not due to semantic priming via form resemblance, as, say, the nonword *floable* might make a person think of roundness because it rhymes with *global*).

Analysis

Because our analyses were numerous and identical for each dimension we considered, we begin here with an outline of the structure of those analyses.

Each pole of each dimension was modeled separately (recall that participants made decisions about only one end of the continuum at a time). For each pole, we computed three separate regression models based on counting the occurrences in each string of three formal characteristics of each nonword string: phonetic features (as defined in International Phonetic Association, 1999)[7]; phonemes (as defined by Apple's phoneme set) and common biphones (defined as those that occurred at least 200 times in our stimulus set); and letters and common bigrams (defined the same way). We used binomial regression models to model the acceptance rates in each category. Predictors were entered together, and removed in order of decreasing p value until only predictors that contributed at a probability of $p < .001$ (usually much less) remained. The t and p values for each predictor are reported.

We also report a fourth hybrid model for each pole of each dimension that used the predictors that had entered into all six base models to predict responses to each pole, removing predictors in the same way until those that remained contributed with $p < .001$. When predictors in the model were perfectly correlated (e.g., letter.l and LATERAL.APPROXIMANT, letter.r and ALVEOLAR. APPROXIMANT), we left in features over phonemes, and phonemes over letters.

For all four models of each dimensional anchor, we report six performance measures that are relevant to assessing how well they performed. The first measure is *the cross-validated hit rate*, the number of times the binary model's classification probability (rounded to a binary integer, i.e., 1 or 0) agreed with the classification of experimental participants for all strings that were actually seen by participants, cross-validated with k-fold cross-validation ($k = 10$). The second measure is *the observed hit rate*, the exact observed hit rate for all strings that were actually seen by participants. These measures are the same, with and without cross-validation; we provide both to confirm that observed hit rates are not due to over-fitting. The observed hit rate is broken down into two parts: *the true positive rate* (how often the model correctly predicted the acceptance of a presented string) and *the true negative rate* (how often the model correctly predicted the rejection of a presented string). For each of these two hit rates we also report d′, a standardized measure of signal detection accuracy. This is necessary because it is possible to achieve a high true positive or true negative rate at the expense of also having a high false positive or false negative rate. In the extreme, it would be possible for a model to predict string acceptance in any category with 100% accuracy simply by predicting that every string was accepted.

These models are all limited in two ways. One way is that they each predict only a single pole of each dimension. It is obvious that models which synthesize the evidence against or in support of membership in the categories defined by both poles are likely to outperform those which predict only a single pole, since the synthetic models are able to access more relevant information. The

[7] Following the IPA consonant chart in this book, we placed /l/ and /r/ into their own categories, rather than breaking them down into lateral and alveolar approximants. The phoneme /r/ is complex and highly variable in English; other decisions might be deemed more appropriate by some.

second limitation is that the models report performance for *all* presented strings. This must be an under-estimation of the true size of any sound symbolism effects since many strings are likely to be completely unrelated to the dimension of interest, and must therefore be judged randomly by the participants in our experiment. By analogy, if we forced experimental participants to decide if presented images represented plants or animals, but included pictures of rocks, we would have to expect random performance on the images of rocks. Participants faced with the unreasonable demand to classify rocks as either animals or plants have no choice but to choose randomly. That some (indeed, most) nonword strings will be non-sound-symbolic within any given category is an inevitable consequence of having only a small number of predictors relevant to any category. Since cues are rare and strings are short (or, to be more precise, since the ratio of the number of available cues to the number of cues per string is high), many strings will certainly contain none of those predictors, and thus contain no relevant discriminative information at all about their suitability to symbolize that category. This is especially true if, as we proposed earlier and will evaluate later, common cues (such as phonological features) are less likely to be discriminative than rare cues.

In order to address these two limitations (i.e., predicting only a single pole of each dimension and reporting performance for all presented strings), we also report performance on the subset of presented strings for which each model had 'high confidence', defined here (in the absence of any formal arguments for or against any difference) as an absolute difference in classification probability > .30. For this subset of strings (and for each pole), we report the last five performance measures described above, as well as the proportion of strings seen by participants that met this criterion. When we wish to reference a model for the entire dimension rather than as model for one pole of the dimension, we will be referencing these hybrid models, using the notation *x/y* to refer to the model defined by subtracting estimates from the high confidence model of *y* from estimates from the high confidence model of *x* (i.e., *masculine/feminine* = estimated probability of being *masculine* - estimated probability of being *feminine*, confined to that subset of strings for which the absolute value of this difference > 0.30).

Having so many measures for so many models of each of two poles of a dimension is a mixed blessing. Although the plethora of measures enables us to assess each model in a fine-grained way, it also complicates any attempt to adjudicate the question: *Which model is best?* A model might have a very high hit rate, but achieve that hit rate by having a very high true negative rate coupled with a very low true positive rate. Or again, a 'high confidence' model might have a high true and false positive rate, but achieve that hit rate by being able to classify only a very tiny proportion of presented strings. Which model is 'best' is ultimately a matter of taste, and a function of what one is trying to achieve and what values one wishes to maximize. We will highlight strengths and weaknesses of each model in our discussion.

For the hybrid models, we present the ten strings predicted by the model to be most and least likely exemplars at each pole, regardless of whether any of those strings were actually judged by any experimental participant. This allows the reader to assess the face validity of the model.

Results

Commonly-studied sound symbolism categories

Because they were selected for different reasons, and modeled using different methods, we present and discuss results from the commonly-studied sound symbolism categories separately from the semantic categories. We discuss the overall statistical properties of the predictors in the general discussion after presenting both categories. We begin with the former, which were modeled by directly asking participants if a string was a good exemplar of a name for a thing described by one of the poles: i.e., *a round thing*.

Dimension 1: Round/Sharp

Perhaps the best-known dimension in sound symbolism research is the dimension of *round/sharp*, first described by Köhler (1929, 1947). Ever since Köhler's work, sharpness has been associated with voiceless stop consonants such as /k/ and /t/, and roundness with sonorants such as /m/ and /l/, as well as with the voiced stop consonant /b/. Some have also suggested that there are vowel predictors of these categories, with sharpness being associated with unrounded front vowels, and roundness being associated with rounded back vowels (e.g., Nielsen & Rendall, 2011). D'Onofrio (2014) extended this, pointing out that the strings originally used by Kohler "differed from one another in vowel roundedness and vowel backness, but also in continuant nature of consonants, sonority of consonants, voicing of consonants, and place of articulation of consonants" (p. 368–369).

The results shown graphically in Fig. 1 and presented numerically in Table 3 suggest that both poles of this dimension are among the most predictable of those we examined, with a maximal classification agreement rate for the category *sharp* among the strings classified with high confidence by the hybrid model as *sharp* of 73% (p < 2E−16; True negative = 43%; d′ = 0.87; True positive = 30%; d′ = 0.62; 31% of seen strings classified) and a maximal classification agreement rate for the category *round* among the strings classified with high confidence using the hybrid model of 73% (p < 2E−16; True negative = 31%; d′ = 0.46; True positive = 43 %; d′ = 1.08; 34% of seen strings classified).

The two hybrid models largely support traditional findings, showing voiceless stop consonants (letter.k, letter.c, letter.t, and letter.x) as indicators of sharpness (and/or negative indicators of roundness) and the voiced stop consonant phoneme.b /b/ as one indicator of roundness (or, more precisely, as a negatively-weighted indicator of sharpness). Three phonemes (central mid phoneme.OW /əʊ/, closed back phoneme.UW /uː/, and open back phoneme.AA /ɒ/) and the continuant letter.m were also weighted as negative signs of sharpness.

The feature models were the weakest, as predicted by a signal detection perspective on sound symbolism. Although they achieved respectable hit rates of 74% for both the *sharp* and the *round* category on the strings classified with high confidence, the feature models did so at a high cost in terms of the percentage of strings they could classify with such confidence (*Sharp*: 11%, *Round*: 12%, compared to 32% respectively for the two classes classified by the hybrid model).

The base models built on phonemes and letters included many expected predictors weighted directionally according to traditional understanding, as discussed above: e.g., positive weights for *sharp* on letter.x, letter.k, letter.c, and letter.t; negative weights for *sharp* on phoneme.b and phoneme.m; positive weights for *round* on letter.o and letter.u; and negative weights for roundness on letter.k, letter.t, phoneme.t /t/, and phoneme.k /k/, as well as on two bigrams, bigram.am and bigram.bl. As shown in Table 4, the letter model makes errors because it does not 'know' that *initial-k* is silent before *n* (as in *knitsky*), or that *initial-x* (as in *xykipt*) is likely to be pronounced as a voiced alveolar fricative /z/. Of course it would be very easy to adjust the models to account for these anomalies, but since they affect only a few words, there are many analogous anomalies in English that are equally deserving of special treatment, and because such post hoc tinkering is against the 'no selection of stimuli' spirit that animates this investigation, we have not attempted to do so.

Table 3
Model summary for categories *sharp* and *round*. Predictors are ordered by decreasing magnitude of beta weight. Predictors with a positive weight are shown in bold. **CV**: K-fold cross-validated hit rate (k = 10). **Hits**: Exact observed hit rate. **TP**: True positive rate. **TN**: True negative rate. **Difference models** are limited to high confidence strings, defined as estimated to have a difference in probability of belonging to one category (*sharp* or *round*) – probability of belonging the other pole > 0.30. **Proportion**: Proportion of seen strings in the difference model.

Model	Sharp	Estimate	SE	t	Pr(>|t|)	Performance	Difference model	Round	Estimate	SE	t	Pr(>|t|)	Performance	Difference model
Features	**(Intercept)**	**0.02**	**0.08**	0.24	0.81	CV: 0.588	PROPORTION: 0.11	**(Intercept)**	**0.61**	**0.12**	**4.97**	**7.19E−07**	CV: 0.608	PROPORTION: 0.12
	BILABIAL	−0.46	0.07	−6.21	6.48E−10	HITS: 0.6	HITS: 0.74	**LATERAL.APPROXIMANT**	**0.54**	**0.09**	**5.88**	**4.67E−09**	HITS: 0.61	HITS: 0.74
	VOICED	−0.3	0.06	−4.62	4.14E−06	TP: 0.17	TP: 0.21	LABIODENTAL	−0.45	0.11	−4.28	1.98E−05	TP: 0.24	TP: 0.53
	VELAR	**0.27**	**0.07**	**3.8**	0.000149	TP d′: −0.29	TP d′: 0.11	VELAR	−0.39	0.07	−5.36	9.14E−08	TP d′: 0.06	TP d′: 1.53
						TN: 0.42	TN: 0.53	FRONT	−0.38	0.06	−6	2.39E−09	TN: 0.37	TN: 0.21
						TN d′: 0.86	TN d′: 1.5	ALVEOLAR	−0.33	0.05	−6.17	8.40E−10	TN d′: 0.65	TN d′: 0.07
								NASAL	**0.31**	**0.07**	**4.1**	4.36E−05		
Phonemes	(Intercept)	−0.07	0.07	−1.01	0.31	CV: 0.617	PROPORTION: 0.27	**(Intercept)**	**0.17**	**0.08**	**2.22**	**0.03**	CV: 0.621	PROPORTION: 0.27
	Phoneme.OW	−0.85	0.19	−4.4	1.15E−05	HITS: 0.62	HITS: 0.69	**Biphone.IYAX**	**1.17**	**0.25**	**4.72**	**2.57E−06**	HITS: 0.62	HITS: 0.7
	Phoneme.UW	−0.82	0.19	−4.41	1.09E−05	TP: 0.14	TP: 0.27	Phoneme.z	−0.81	0.21	−3.77	1.65E−04	TP: 0.28	TP: 0.4
	Phoneme.b	−0.72	0.12	−5.87	5.21E−09	TP d′: −0.54	TP d′: 0.39	Phoneme.t	−0.54	0.1	−5.45	5.51E−08	TP d′: 0.3	TP d′: 1.01
	Phoneme.k	**0.68**	**0.09**	**7.59**	4.86E−14	TN: 0.48	TN: 0.42	Phoneme.IY	−0.53	0.12	−4.26	2.15E−05	TN: 0.35	TN: 0.3
	Phoneme.m	−0.59	0.12	−5.04	5.05E−07	TN d′: 1.33	TN d′: 0.83	Phoneme.k	−0.46	0.09	−4.89	1.11E−06	TN d′: 0.49	TN d′: 0.35
	Phoneme.AA	−0.46	0.12	−3.9	0.0001			Phoneme.IH	−0.44	0.1	−4.48	7.98E−06		
								Phoneme.m	**0.44**	**0.11**	**3.92**	**8.97E−05**		
								Phoneme.b	**0.42**	**0.11**	**3.79**	**0.000154**		
Letters	(Intercept)	−0.37	0.08	−4.53	6.13E−06	CV: 0.624	PROPORTION: 0.31	(Intercept)	−0.35	0.07	−4.74	2.27E−06	CV: 0.636	PROPORTION: 0.33
	Letter.x	**1.87**	**0.47**	**3.96**	**7.89E−05**	HITS: 0.63	HITS: 0.71	Letter.k	−1.04	0.17	−6.12	1.11E−09	HITS: 0.64	HITS: 0.71
	Letter.k	**0.75**	**0.15**	**5.05**	**4.84E−07**	TP: 0.22	TP: 0.29	**Bigram.am**	**0.96**	**0.25**	**3.85**	**0.00012**	TP: 0.26	TP: 0.41
	Letter.c	**0.57**	**0.09**	**6.44**	**1.47E−10**	TP d′: 0.04	TP d′: 0.53	**Bigram.bl**	**0.86**	**0.23**	**3.73**	**0.00020**	TP d′: 0.2	TP d′: 1.04
	Letter.b	−0.53	0.11	−4.73	2.42E−06	TN: 0.41	TN: 0.41	**Letter.o**	**0.63**	**0.08**	**8.19**	**4.63E−16**	TN: 0.37	TN: 0.31
	Letter.m	−0.51	0.12	−4.39	1.18E−05	TN d′: 0.8	TN d′: 0.8	Letter.t	−0.47	0.09	−5.54	3.45E−08	TN d′: 0.67	TN d′: 0.37
	Letter.t	**0.39**	**0.08**	**4.82**	**1.53E−06**			**Letter.u**	**0.47**	**0.1**	**4.83**	**1.46E−06**		
Composite	(Intercept)	−0.33	0.08	−4.05	5.25E−05	CV: 0.636	PROPORTION: 0.31	(Intercept)	−0.27	0.09	−3.0	0	CV: 0.648	PROPORTION: 0.34
	Letter.x	**1.81**	**0.47**	**3.83**	**0.000134**	HITS: 0.64	HITS: 0.73	**Bigram.am**	**1.04**	**0.25**	**4.13**	**3.73E−05**	HITS: 0.65	HITS: 0.73
	Letter.k	**0.78**	**0.15**	**5.18**	**2.45E−07**	TP: 0.22	TP: 0.3	Letter.k	−0.74	0.19	−3.86	0.000116	TP: 0.27	TP: 0.43
	Phoneme.OW	−0.77	0.19	−3.98	7.00E−05	TP d′: 0.03	TP d′: 0.62	Letter.o	0.66	0.08	8.43	<2E−16	TP d′: 0.22	TP d′: 1.08
	Phoneme.UW	−0.77	0.19	−4.1	4.32E−05	TN: 0.42	TN: 0.43	LABIODENTAL	−0.48	0.11	−4.52	6.65E−06	TN: 0.38	TN: 0.31
	Phoneme.b	−0.62	0.12	−5.01	5.93E−07	TN d′: 0.84	TN d′: 0.87	Letter.t	−0.48	0.09	−5.47	4.95E−08	TN d′: 0.72	TN d′: 0.46
	Letter.c	**0.58**	**0.09**	**6.41**	1.84E−10			**Letter.u**	**0.47**	**0.1**	**4.78**	**1.84E−06**		
	Letter.m	−0.52	0.12	−4.51	6.77E−06			k	−0.41	0.11	−3.8	1.49E−04		
	Phoneme.AA	−0.45	0.12	−3.8	0.000147			**LATERAL.APPROXIMANT**	**0.34**	**0.08**	**4**	**6.48E−05**		
	Letter.t	**0.41**	**0.08**	**5.03**	5.20E−07									

Table 4
Ten strings predicted to be highest and lowest in probability of belonging to the categories of *sharp* and *round*, from the 7996 strings used in the experiment.

Category	Sharp	p(Sharp)	Round	p(Round)	Sharp – Round	p(Sharp) – p(Round)
High	axittic	0.95	amorul	0.9	knitick	0.83
High	xykipt	0.94	hoonous	0.9	axittic	0.83
High	exiduct	0.92	hoorous	0.9	xykipt	0.8
High	fixtant	0.91	boamion	0.89	fixtant	0.8
High	restatex	0.91	broodam	0.89	crickty	0.79
High	exignak	0.91	hoonsam	0.89	keppick	0.79
High	karx	0.91	coulous	0.87	kanktil	0.77
High	keex	0.91	ambous	0.87	knitsky	0.77
High	knitick	0.9	hambous	0.87	karx	0.77
High	cruckwic	0.9	amious	0.87	keex	0.77
Low	bomble	0.07	anktify	0.09	ambula	−0.73
Low	bombal	0.07	keefify	0.09	goomon	−0.73
Low	butomy	0.07	cafttic	0.08	buroong	−0.74
Low	balmo	0.06	keppick	0.07	eposomo	−0.76
Low	bobluay	0.06	krenker	0.07	honulo	−0.78
Low	dobulum	0.05	kark	0.07	broodam	−0.79
Low	boomeo	0.05	kuket	0.07	boamion	−0.79
Low	boodoma	0.05	knitick	0.07	dobulum	−0.79
Low	brimbom	0.04	knitsky	0.07	boomeo	−0.8
Low	bugovo	0.04	kanktil	0.06	boodoma	−0.8

Table 5
Summed sharpness weight for predictors that appeared in the hybrid model for *sharp/round*.

Predictor	Summed sharpness β
Letter.x	1.81
Letter.k	1.52
Letter.t	0.89
Letter.c	0.58
LABIODENTAL	0.48
Phoneme.k	0.41
LATERAL.APPROXIMANT	−0.34
Phoneme.AA	−0.45
Letter.u	−0.47
Letter.m	−0.52
Phoneme.b	−0.62
Letter.o	−0.66
Phoneme.OW	−0.77
Phoneme.UW	−0.77
Bigram.am	−1.04

Table 4 shows the top ten nonwords from the 7996 in the experiment that the hybrid model would judge as either *sharp* or *round*, with the probability assigned to them by regression. By subtracting the probability estimate for one category from the other (essentially creating a composite model of the dimension by summing one model with a negatively-weighted version of the other) we can get an overall weight on the dimension. The top and bottom nonwords on that model are also shown in Table 4. Across all 7996 strings, the estimates for the two poles are negatively correlated at $r = -0.66$ ($p < 2E-16$; see Fig. 3), a highly reliable correlation that is nevertheless perhaps lower might be expected for two dimensions defined as opposites.

Discussion. As we suggested in the introduction, one benefit of using regression to identify sound symbolic features is that it not only allows those features to be *identified*, but also allows them to be *weighted*. This makes it possible to address with some quantitative precision questions in the literature about whether it is vowels or consonants that primarily drive the *round/sharp* sound symbolism effect. To get the true weight for any character that appears with inverse weights in both the *round* and *sharp* model, we need to sum their absolute values: e.g., letter.k appears in the hybrid round model with a β of −0.74 and in the sharp model with a β of 0.78, for a total β *sharp* weight of 1.52. The results of this exercise are shown in Table 5. The strongest predictor is letter.x (Total β *sharp* weight: 1.81) followed closely by letter.k (Total β *sharp* weight: 1.52). The third strongest predictor, with 59% the weight of the second, is letter.t (Total β *sharp* weight: 0.89). The three vowels (Phoneme.AA, Phoneme.OW and Phoneme.UW) have total β *sharp* weights of −0.45, −0.77 and −0.77 respectively, suggesting that they each carry less than 42% of the weight of Letter.x (although this is partially mitigated by the separate presence of Letter.o, with a sharpness weight of −0.66). This finding is consistent with the theoretical considerations in the introduction, from which it was predicted that common cues must necessarily be weak cues. Letter.m is a much weaker predictor than the stop consonants, with a total β *sharp* weight of −0.52, just under a third of the weight on letter/phoneme k. More generally, the models make clear that the stop consonants carry a very large proportion of the weight in accounting for sound symbolism in this category.

One finding not previewed in the literature is the strong weighting on *round* for Bigram.am. It has a large total β *sharp* weight of −1.56 (−1.04 itself, plus the weight of letter.m, an additional −0.52), about 86% of the weight of letter.k, but in the opposite direction. As a result of this strong weight, the bigram appears in several of the strings most strongly predicted to be *round* such as *amorul, boamion,* and *broodam*. Strings containing the Bigram.am were accepted as *round* 66.25% of the time they were seen by experimental participants ($n = 80$), compared to being accepted as *sharp* just 40% of the time they were seen ($n = 65$), a reliable difference ($\chi^2(1)$ with Yates' correction = 6.42, two-tailed $p = .01$).

One question we would like to be able to address is: *Why* is phoneme.b, a voiced plosive, associated with *round* (as previous work and the models presented here suggest) when other voiced plosives (letter.d and letter.g) are normally not mentioned as being associated with either *sharp* or *round*, and do not appear in the letter model? Although the models do weight phoneme.b /b/ with the category *round* (Total β *sharp* weight: −0.62), they do not provide any clear answer to the question of why /d/ and /g/ do not, since none of the features that distinguish the three plosives appears in any of the models. We can however look at this question from the perspective of signal detection principles, and will return to this in the final discussion, when we will have relevant evidence from other models.

How would these models adjudicate the most famous *sharp/round* stimuli, *takete* vs. *maluma*? The models assign *takete* a 61.1% chance of being *sharp* and an 8.5% chance of being *round*, for a difference of 52.6%, well past the high certainty threshold we have set. It is more difficult to adjudicate *maluma*, since it depends on how it is pronounced. Assuming a short u /ʌ/, the string

Table 6

Model summary for categories *large* and *small*. Predictors are ordered by decreasing magnitude of beta weight. Predictors with a positive weight are shown in bold. **CV**: K-fold cross-validated hit rate (k = 10). **Hits**: Exact observed hit rate. **TP**: True positive rate. **TN**: True negative rate. **Difference models** are limited to high confidence strings, defined as estimated to have a difference in probability of belonging to one category (*large* or *small*) – probability of belonging the other pole > 0.30. **Proportion**: Proportion of seen strings in the difference model.

Model	Large	Estimate	SE	t	Pr(>|t|)	Performance	Difference model	Small	Estimate	SE	t	Pr(>|t|)	Performance	Difference model
Features	(Intercept)	−0.94	0.1	−9	<2E−16	CV: 0.58	PROPORTION: 0.19	**(Intercept)**	**0.67**	**0.11**	**6.18**	**7.7E−10**	CV: 0.58	PROPORTION: 0.21
	BACK	**0.36**	**0.07**	**4.79**	**1.83E−06**	HITS: 0.58	HITS: 0.68	POSTALVEOLAR	−0.77	0.16	−4.98	0.000000699	HITS: 0.58	HITS: 0.64
	VOICED	**0.2**	**0.04**	**4.88**	**1.16E−06**	TP: 0.01	TP: 0.01	GLOTTAL	−0.77	0.19	−4.1	0.0000421	TP: 0.26	TP: 0.61
						TP d′: −2.11	TP d′: −1.84	VOICED	−0.3	0.04	−7.04	2.58E−12	TP d′: 0.17	TP d′: 2.64
						TN: 0.57	TN: 0.67						TN: 0.31	TN: 0.03
						TN d′: 2.67	TN d′: 2.62						TN d′: 0.31	TN d′: −1.55
Phonemes	(Intercept)	−0.77	0.06	−12.41	<2E−16	CV: 0.626	PROPORTION: 0.013	**(Intercept)**	**0.08**	**0.06**	**1.42**	**0.16**	CV: 0.55	PROPORTION: 0.0164
	biphone.AXs	**0.89**	**0.22**	**4.15**	**3.51E−05**	HITS: 0.63	HITS: 0.74	Phoneme.r	−0.4	0.08	−5.33	1.11E−07	HITS: 0.55	HITS: 0.63
	g	**0.87**	**0.14**	**6.39**	**1.99E−10**	TP: 0.14	TP: 0.74						TP: 0.3	TP: 0
	AA	**0.64**	**0.12**	**5.47**	**4.98E−08**	TP d′: −0.53	N/A						TP d′: 0.44	N/A
	b	**0.58**	**0.11**	**5.52**	**3.91E−08**	TN: 0.49	TN: 0						TN: 0.25	TN: 0.63
	biphone.AXn	**0.53**	**0.14**	**3.83**	**0.000133**	TN d′: 1.32	N/A						TN d′: −0.09	N/A
Letters	(Intercept)	−0.98	0.08	−11.75	<2E−16	CV: 0.62	PROPORTION: 0.023	**(Intercept)**	**0.25**	**0.06**	**3.94**	**8.41E−05**	CV: 0.58	PROPORTION: 0.0294
	Bigram.oo	−1.05	0.28	−3.75	0.00018	HITS: 0.63	HITS: 0.66	Letter.h	−0.49	0.11	−4.51	6.93E−06	HITS: 0.58	HITS: 0.65
	Bigram.um	**0.96**	**0.25**	**3.87**	**0.00011**	TP: 0.13	TP: 0.66	Letter.g	−0.43	0.11	−3.98	7.04E−05	TP: 0.22	TP: 0
	Letter.o	**0.69**	**0.09**	**7.55**	**6.38E−14**	TP d′: −0.55	N/A	Letter.r	−0.4	0.07	−5.39	7.75E−08	TP d′: −0.09	N/A
	Letter.g	**0.66**	**0.11**	**5.94**	**3.28E−09**	TN: 0.5	TN: 0						TN: 0.36	TN: 0.65
	Letter.b	**0.48**	**0.1**	**4.87**	**1.22E−06**	TN d′: 1.4	N/A						TN d′: 0.6	N/A
	Letter.a	**0.27**	**0.07**	**3.78**	**0.000159**									
Composite	(Intercept)	−0.87	0.07	−12.36	<2E−16	CV: 0.614	PROPORTION: 0.14	**(Intercept)**	**0.69**	**0.11**	**6.19**	**7.11E−10**	CV: 0.58	PROPORTION: 0.15
	Bigram.oo	−1.06	0.28	−3.78	0.00016	HITS: 0.62	HITS: 0.73	Letter.h	−0.66	0.11	−5.76	9.87E−09	HITS: 0.5	HITS: 0.63
	Phoneme.g	**0.88**	**0.14**	**6.48**	**1.14E−10**	TP: 0.13	TP: 0.01	VOICED	−0.31	0.04	−7.16	1.15E−12	TP: 0.44	TP: 0.61
	Phoneme.b	**0.52**	**0.11**	**4.88**	**1.13E−06**	TP d′: −0.55	TP d′: −1.59						TP d′: 1.76	TP d′: 2.76
	Letter.o	**0.5**	**0.09**	**5.67**	**1.63E−08**	TN: 0.49	TN: 0.72						TN: 0.06	TN: 0.03
	BACK	**0.33**	**0.08**	**4.21**	**2.66E−05**	TN d′: 1.3	TN d′: 3.07						TN d′: −1.49	TN d′: −1.61

Table 7
Ten strings predicted to be highest and lowest in probability of belonging to the categories of *large* and *small*, from the 7996 strings used in the experiment.

Large	p(Large)	Small	p(Small)	Large – Small	p(Large) – p(Small)
globlor	0.9	accrel	0.67	glozzho	0.44
sogung	0.89	aciess	0.67	bomburg	0.43
bomburg	0.87	ackel	0.67	bugovo	0.43
bugovo	0.87	acken	0.67	gragwom	0.41
globion	0.87	acket	0.67	bongard	0.41
globson	0.87	ackey	0.67	globlor	0.38
grobson	0.87	ackic	0.67	sogung	0.37
canagog	0.85	ackie	0.67	horgous	0.36
epsigog	0.85	ackiff	0.67	gotod	0.36
glonk	0.85	ackis	0.67	globion	0.35
ackiff	0.3	chithet	0.35	ackis	−0.37
ackie	0.3	chirish	0.35	ackiff	−0.37
ackic	0.3	chenth	0.35	ackie	−0.37
ackey	0.3	chash	0.35	ackic	−0.37
acket	0.3	chalish	0.35	ackey	−0.37
acken	0.3	chalchin	0.35	acket	−0.37
ackel	0.3	hurgh	0.28	acken	−0.37
aciess	0.3	hanzhor	0.28	ackel	−0.37
achem	0.3	hatcherb	0.28	aciess	−0.37
accrel	0.3	chithway	0.28	accrel	−0.37

has a 54.9% chance of being *round* and a 20.2% chance of being *sharp*, for a difference of 34.7%, again past the high confidence threshold for being *round*. If it is rather a long /uː/, there is an additional piece of strong evidence against the string being sharp, dropping the *sharp* probability to just 10.5%.

Dimension 2: Large/Small

Perhaps the second most common dimension for sound symbolism in the previous literature is *large/small*. In general, research has suggested that nonwords containing open-back vowels/closed-front vowels are more appropriate as labels for large and small targets, respectively (Newman, 1933; Sapir, 1929). There is also some evidence that this association has affected existing vocabularies. Jespersen (1925) noted long ago that:

> "The vowel [i], especially in its narrow or thin variety, is particularly appropriate to express what is small, weak, insignificant, or, on the other hand, refined or dainty. It is found in a great many adjectives in various languages, e.g., *little, petit, piccolo, piccino*, Magy. [Magyar, or Hungarian], *kis*, E. *wee, tiny*, (by children often produced *teeny* [...]), *slim*, Lat. *minor, minimus*, Gr. *mikros* [...] The same vowel is found in diminutive suffixes in a variety of languages as E. *–y, -ie* (*Bobby, baby, auntie, birdie*), Du. *–ie, -je* (*koppie* 'little hill'), Gr. -i- (*paid-i-on* 'little boy'), Goth. *–ein* [...] (*gumein* 'little man'), E. *–kin, -ling*, Swiss German *–li*, It. *–ino*, Sp. *–ico, -ito, -illo*...[*sic*]" (p. 402).

Ultan (1978) surveyed 136 languages and found that those using vowel ablauting to express diminutive concepts tended to do so with closed-front vowels. In addition, Blasi et al. (2016) found that in their sample of nearly two-thirds of the world's languages, words for small tended to contain the vowel /i/.

The models for these two classes are presented in Table 6. They differ in several ways from the models for *sharp/round*, notably by being generally worse models. Also notable is the fact that there are many positively weighted predictors for *large*, but none (and relatively few negative predictors) for *small*.

The high-confidence phonological feature model for *large* achieves a large negative true positive d′ value of −2.11. As this value suggests, the model attains its hit-rate of 68% (on 19% of seen stimuli) almost entirely on the strength of true negatives (True negative rate: 0.67; True negative d′: 2.62), at the expense of true positives. The high confidence phonological feature model for *small* classified 21% of seen stimuli, with a true positive rate of 0.61 (d′ = 2.64) but a poor true negative rate of 0.03 (d′ = −1.55).

The phoneme model for *small* is worse, consisting of just a single negatively-weighted predictor (β = −0.4), phoneme.r. Not surprisingly given the paucity of predictors, the high confidence model was poor, with a hit rate of 63%, attained entirely by true negatives. Moreover, it only classified 1.6% of all judged words, indicating that it differentiated little in probability of classifying a string to either category. The letter model was about as poor, and can be characterized sufficiently here by noting that the high confidence model also had a true positive rate of 0.

The hybrid model for *large* consisted of four positively weighted cues (phoneme.g, phoneme.b, letter.o, and the feature BACK, all consistent with previous work) and one strongly negatively weighted feature, bigram.oo (β = −1.06). This high confidence model was able to correctly classify 73% of 14% of seen stimuli, though again almost entirely on the strength of correctly identifying true negatives (True positive rate: 0.01, d′ = −1.59; True negative rate = 0.72, d′ = 3.07). The high confidence hybrid model for *small* consisted of just two weakly negative predictors: letter.h (β = −0.66) and the feature VOICED (β = −0.31). Note that these two predictors are both common and have low β weights, as predicted by a signal detection view of sound symbolism. The model nevertheless achieved a good true positive rate (0.61, d′ = 2.76) in classifying the 15% of seen stimuli that showed a strong difference in classification probability (True negative rate = 0.03, d′ = −1.61).

The 20 strings judged by the hybrid models to be most strongly weighted in either category are shown in Table 7. Note that the probabilities attached to words judged *large* are much higher (with a maximum p = 0.90) than the probabilities attached to words judged *small* (maximum p = 0.67). The estimates are negatively correlated at just *r* = −0.34 (*p* < 2E−16), suggesting a large degree of independence between these two poles. As shown in Fig. 3, the estimates from *small* are almost uncorrelated with any other measures. This contrasts with the estimates for *large*, which shows a strong positive correlation with estimates for the category *round* (*r* = 0.47, *p* < 2E−16); strong negative correlations with estimates for the categories *feminine* (*r* = −0.39, *p* < 2E−16), *sharp* (*r* = −0.31, *p* < 2E−16) and *flower* (*r* = −0.28, *p* < 2E−16); and a highly reliable but weaker positive correlation with the category *masculine* (*r* = 0.09, *p* < 2E−16). This seems to suggest a central role for *large* in sound symbolism, since sound symbolism in many other categories seems to partially 'piggyback' on it.

Discussion. In this case, it might be argued that the hybrid models are not the best models. The phonological feature models classify a larger proportion of stimuli almost as well or better than the

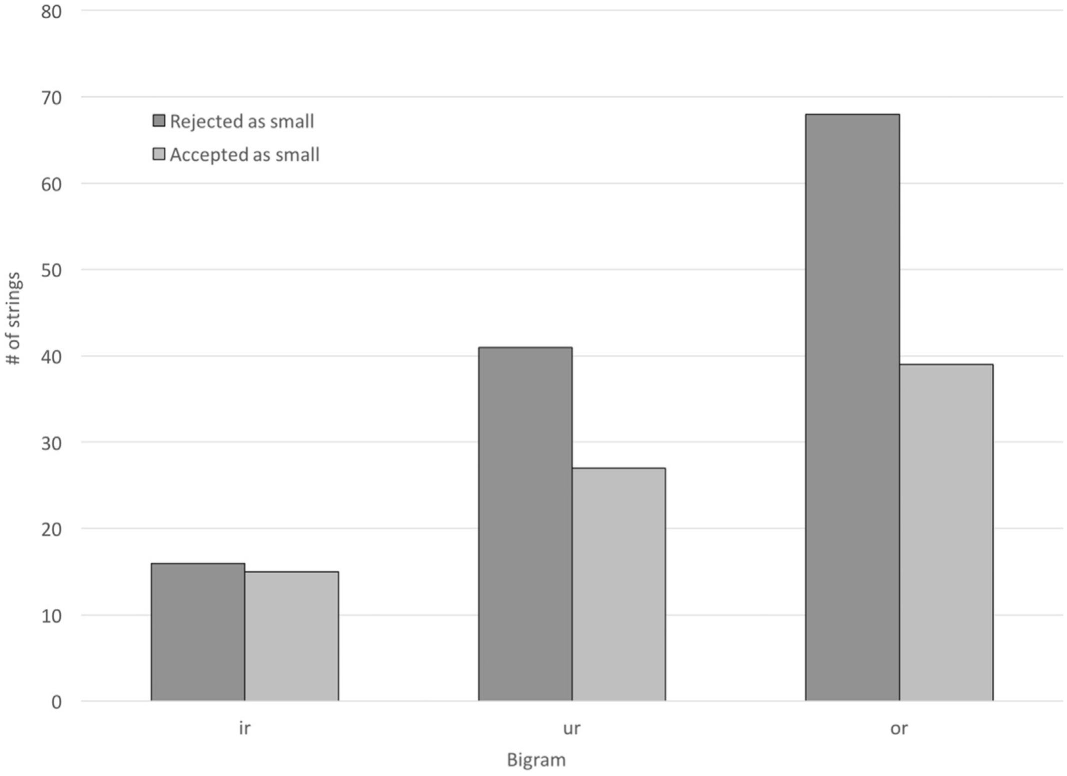

Fig. 2. Acceptance and rejection rates as *small*, for strings containing vowel-r.

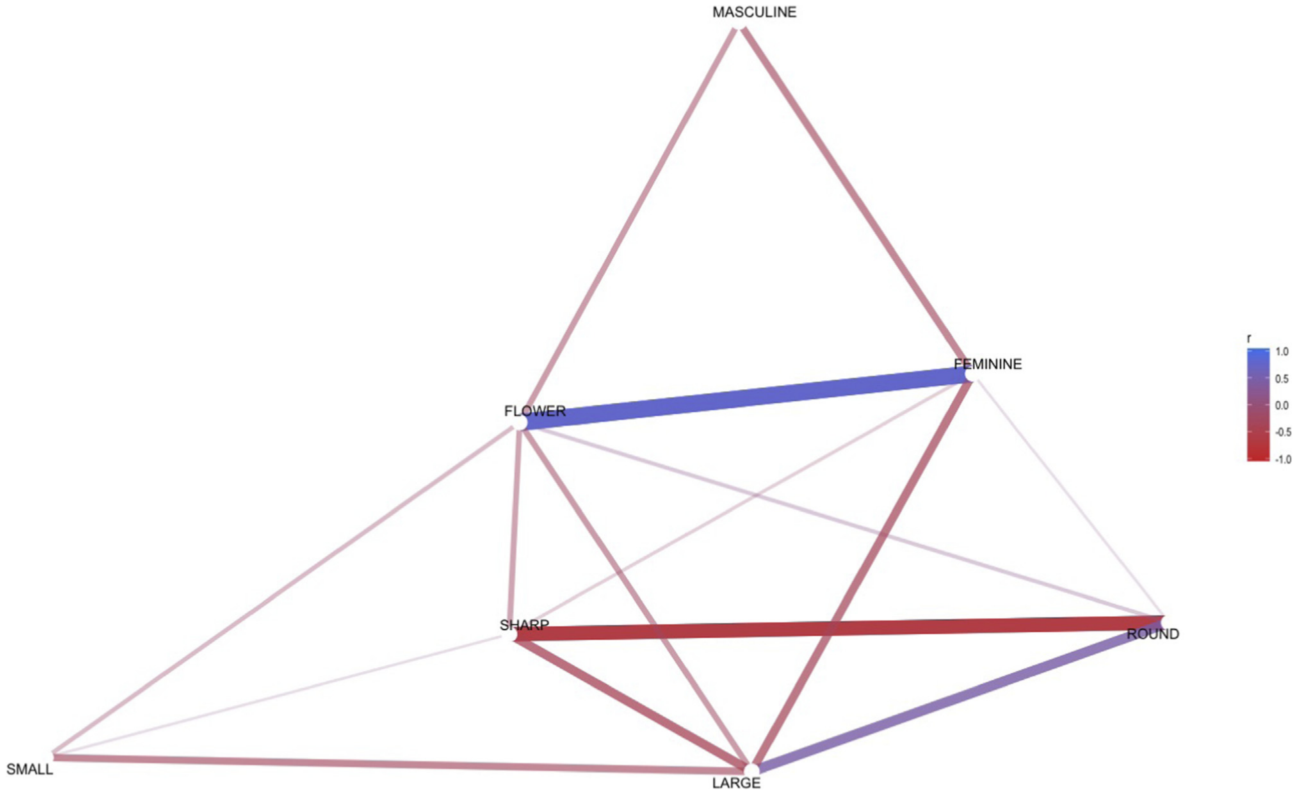

Fig. 3. Graphic representation of correlations between hybrid model estimates. Link width and color both show correlation magnitude. Correlations with a magnitude less than 0.1 are not shown.

hybrid models, with the same pattern of results (excellent true negative performance for *large*; and excellent true positive performance for *small*) and using only four predictors: BACK, POSTALVEOLAR, GLOTTAL, and VOICED (which appears in both models with opposite signs). These models can be succinctly (if somewhat over simplistically) summarized by saying that strings that do not

Table 8
Model summary for categories *masculine* and *feminine*. Predictors are ordered by decreasing magnitude of beta weight. Predictors with a positive weight are shown in bold. **CV**: K-fold cross-validated hit rate (k = 10). **Hits**: Exact observed hit rate. **TP**: True positive rate. **TN**: True negative rate. **Difference models** are limited to high confidence strings, defined as estimated to have a difference in probability of belonging to one category (*masculine* or *feminine*) – probability of belonging the other pole > 0.30. **Proportion**: Proportion of seen strings in the difference model.

Model	Masculine	Estimate	SE	t	Pr(>|t|)	Performance	Difference model	Feminine	Estimate	SE	t	Pr(>|t|)	Performance	Difference model
Features	**(Intercept)**	**0.11**	**0.06**	**2.02**	**0.04**	CV: 0.55	PROPORTION: 0.038	(Intercept)	−0.48	0.14	−3.47	5.29E−04	CV: 0.62	PROPORTION: 0.043
	CLOSE	−0.29	0.06	−4.52	6.68E−06	HITS: 0.55	HITS: 0.63	VELAR	−0.46	0.08	−5.98	2.57E−09	HITS: 0.60	HITS: 0.7
						TP: 0.29	TP: 0.26	**CLOSE.MID**	**0.46**	**0.09**	**5.34**	**1.06E−07**	TP: 0.21	TP: 0.36
						TP d′: 0.28	TP d′: 0.35	**FRONT**	**0.45**	**0.07**	**6.86**	**9.01E−12**	TP d′: 0.09	TP d′: 0.94
						TN: 0.27	TN: 0.37	**LATERAL.APPROXIMANT**	**0.35**	**0.09**	**4.07**	**4.78E−05**	TN: 0.39	TN: 0.34
						TN d′: 0.06	TN d′: 0.48	VOICED	−0.25	0.05	−5.38	8.13E−08	TN d′: 0.52	TN d′: 0.44
Phonemes	**(Intercept)**	**0.08**	**0.05**	**1.59**	**0.11**	CV: 0.54	PROPORTION: 0.0697	(Intercept)	−0.09	0.07	−1.27	0.21	CV: 0.61	PROPORTION: 0.069
	biphone.lIH	−1.08	0.29	−3.77	1.69E−04	HITS: 0.54	HITS: 0.68	Phoneme.g	−−0.73	0.15	−4.74	2.26E−06	HITS: 0.62	HITS: 0.77
	Phoneme.IY	−0.52	0.11	−4.71	2.64E−06	TP: 0.41	TP: 0.61	**Phoneme.f**	**0.54**	**0.12**	**4.63**	**3.87E−06**	TP: 0.05	TP: 0.07
						TP d′: 1.21	TP d′: 2.75	Phoneme.r	−0.47	0.08	−5.76	9.77E−09	TP d′: −1.21	TP d′: −0.66
						TN: 0.13	TN: 0.07	Phoneme.k	−0.4	0.09	−4.31	1.69E−05	TN: 0.56	TN: 0.7
						TN d′: −0.81	TN d′: −1.02						TN d′: 1.96	TN d′: 2.45
Letters	**(Intercept)**	**0.16**	**0.06**	**2.84**	**0.005**	CV: 0.57	PROPORTION: 0.12	(Intercept)	−0.57	0.08	−7.2	8.47E−13	CV: 0.64	PROPORTION: 0.11
	Bigram.fl	−1.25	0.29	−4.32	1.63E−05	HITS: 0.57	HITS: 0.63	Bigram.cr	−1.33	0.37	−3.59	0.000344	HITS: 0.63	HITS: 0.83
	Bigram.ic	**0.9**	**0.22**	**4.01**	**6.27E−05**	TP: 0.32	TP: 0.54	Bigram.ar	−0.8	0.18	−4.47	8.10E−06	TP: 0.1	TP: 0.08
	Letter.y	−0.69	0.17	−4.19	2.88E−05	TP d′: 0.52	TP d′: 2.1	Letter.k	−0.77	0.16	−4.86	1.27E−06	TP d′: −0.77	TP d′: −0.29
	Letter.i	−0.39	0.08	−4.76	2.10E−06	TN: 0.24	TN: 0.09	Letter.g	−0.66	0.12	−5.39	7.72E−08	TN: 0.53	TN: 0.75
						TN d′: −0.08	TN d′: −0.97	**Letter.a**	**0.44**	**0.07**	**5.89**	**4.45E−09**	TN d′: 1.56	TN d′: 2.52
								Letter.i	**0.39**	**0.08**	**4.94**	**8.37E−07**		
Composite	**(Intercept)**	**0.16**	**0.06**	**2.84**	**0.005**	CV: 0.57	PROPORTION: 0.12	(Intercept)	−0.46	0.14	−3.3	0.00097	CV: 0.62	PROPORTION: 0.11
	Bigram.fl	−1.25	0.29	−4.32	1.63E−05	HITS: 0.57	HITS: 0.63	Phoneme.k	−0.59	0.1	−6.12	1.12E−09	HITS: 0.61	HITS: 0.7
	Bigram.ic	**0.9**	**0.22**	**4.01**	**6.27E−05**	TP: 0.32	TP: 0.54	Letter.g	−0.5	0.13	−3.96	7.86E−05	TP: 0.24	TP: 0.34
	Letter.y	−0.69	0.17	−4.19	2.88E−05	TP d′: 0.52	TP d′: 2.1	**CLOSE.MID**	**0.48**	**0.09**	**5.57**	**2.96E−08**	TP d′: 0.27	TP d′: 1.12
	Letter.i	−0.39	0.08	−4.76	2.10E−06	TN: 0.24	TN: 0.09	**FRONT**	**0.47**	**0.07**	**7.18**	**9.63E−13**	TN: 0.37	TN: 0.36
						TN d′: −0.08	TN d′: −0.97	**LATERAL.APPROXIMANT**	**0.35**	**0.09**	**4.13**	**3.83E−05**	TN d′: 0.41	TN d′: 0.35
								VOICED	−0.25	0.05	−5.21	2.06E−07		

contain voiced consonants or back vowels will not be judged *large*, and strings that do contain voiced glottal, voiced, or postalveolar consonants will not be judged *small*. This very brief summary is an over-simplification because the large intercepts in both models add some additional constraints, most notably that strings will need several back vowels and/or voiced consonants before they will be classified by the model as *large*, since the vowels cues are weakly weighted. Since VOICED appears in both models with opposite signs, it has a total β *sharp* weight in the high confidence phonological feature models of 0.5, not just 0.2 as weighted in the model for *large*. The predictor VOICED is a very strong constraint, since there are thirteen voiced consonants.

Taken together, all the models suggest that there is an *a priori* bias against strings being judged as *large* (most success in all *large* models is attributable to success at identifying true negatives) and in favor of them being judged as *small*. Less than a third as many strings are judged with high confidence to belong to either category as were judged to belong to the categories of *round* and *sharp*.

The negative weight of *r* in the letter and phoneme models for *small* strings is a novel finding, though there is no similar predictor in the final hybrid model, perhaps because of the effect of the feature VOICED. It is a reasonable hypothesis that the exclusion of *r* in the model for *small* is to exclude r-controlled vowels (i.e., *ar, er, ir, or, ur, yr*) that are not coded phonologically. However, we did not find support for this hypothesis. Although participants did reject strings containing *r* preceded by a vowel slightly more often (61% of the time) than they rejected other strings containing *r* (59% of the time), the small difference was not reliable ($\chi^2(1) = 1.18$, one-tailed $p > .05$).

It has been noted that "vowels often lower in the environment of *r*" (Lindau, 1978, p. 556) due to the retraction of the tongue root that is required to produce the sound (Delattre, 1956). Since *r* almost always occurs before or after a vowel,[8] it may be that the negative association of *r* with *small* is due to the effect of the lowered vowel, since lower sounds are associated with larger entities (Fitch, 1997; Ohala, 1983, 1984). We were able to test a closely related hypothesis: that words containing *r* followed by a back (lower) vowel should be rejected at a higher rate than words containing *r* followed by a front (higher) vowel. The relevant data are graphed in Fig. 2, which shows that front-vowel strings containing *ir* are accepted about as often as they are rejected, whereas strings containing back-vowel *ur* or *or* are rejected more often than they are accepted. However, by X^2 test, these differences are also not reliable (ir:ur: $\chi^2(1)$ with Yates' correction = 0.35, one-tailed $p > .05$; ir:or: $\chi^2(1)$ with Yates' correction = 0.98, one-tailed $p > .05$).

Table 7 shows that the ten strings judged by the model to be most likely to be large often combine *r* with phoneme.g, that are weighted positively in the hybrid model for *large*, as in the strings *gragwom, grobson*, and *gragula*. This is a predictable consequence of the model since combining *r* with *g* raises the VOICED feature count, pushing the string away from being considered as a candidate for *small* (three voiced consonants are sufficient to constitute negative evidence against *small* in the hybrid model). Strings containing *gr* were accepted as *large* 69.8% (44/63) of the time they were judged by humans. By contrast, strings that did not contain *gr* were accepted 40.7% (817/2005) of the time they were seen. By χ^2 test, this is a reliable difference ($\chi^2(1)$ with Yates' correc tion = 6.78, one-tailed $p = .004$), supporting the idea that the presence in a string of *gr* is associated with acceptance of that string as *large*. The same weighting considerations apply to voice plosive *br*, but there are no strings containing *br* in the lists in Table 7. Strings containing *br* were accepted as *large* 54.7% (40/73) of the time they were judged by humans, not reliably different from the acceptance rate of words that contained neither *gr* nor *br* ($\chi^2(1)$ with Yates' cor rection = 1.88, one-tailed $p = .08$). We will discuss the sound symbolic differences between the voiced plosive consonants (including also the conspicuously absent *d*) in the general discussion.

We may ask the same question as we asked for *sharp/round* above: How does the model perform with perhaps the most famous contrasting *large/small* pair, Sapir's *mil* versus *mal*? Our models do not include any characteristics that distinguish between these strings, although by weighting the feature BACK positively for *large*, they do suggest a better contrast would have been (e.g.,) *mil/mol* (though now we have two English word homophones, and a mill is almost always smaller than a mall) or *mil/mool*.

Dimension 3: Masculine/Feminine

Previous work on masculine/feminine sound symbolism has drawn parallels to work on both sharp/round and large/small sound symbolism. Sidhu and Pexman (2015); see also Sidhu, Pexman, & Saint Aubin, 2016; Cassidy, Kelly, & Sharoni, 1999) showed a relationship between female names and round shapes and between male names and sharp shapes. Sidhu and Pexman (2015) also demonstrated that the most frequent female names contained significantly more consonants that would be considered round-sounding than sharp-sounding. As mentioned briefly in the introduction, Jespersen (1925) specifically suggested that the phonological predictors of smallness should also be predictors of femininity since "smallness and weakness are often taken to be characteristic of the female sex" (p. 402). Indeed, Tarte (1982) found that the closed-front vowel /i/ was associated with femininity, while the open-back vowel /ɑ/ was associated with masculinity.

The models for predicting human acceptability judgments in these two categories are shown in Table 8. They are weaker than the models we have considered so far.

The feature model for *masculine* contains a single negatively-weighted predictor, CLOSE. The high-confidence model classified just 3.8% of seen stimuli. It achieved a hit rate of 63% on that subset, with a true positive rate of 0.26 (d′ = 0.35) and a true negative rate of 0.37 (d′ = 0.48). The feature model for feminine consisted of five predictors, positively-weighted FRONT, CLOSE.MID, and LATERAL.APPROXIMANT and negatively weighted VELAR and VOICED. The high-confidence model achieved an over-all hit rate of 70% on 4.3% of seen strings (True positive rate = 0.36, d′ = 0.94; True negative rate = 0.34, d′ = 0.44). In sum, the feature models were stronger at predicting decisions for *feminine* than *masculine*.

The high-confidence phoneme models of these poles show a different pattern, and were both poor models. The phoneme model for masculine again contained just two negatively-weighted predictors, the CLOSE phoneme IY /iː/, traditionally associated with smallness, and the biphone lIH /lɪ/. These two features allowed it to classify 6.9% of seen strings with high confidence, with a hit rate of 68% achieved at the cost of a large negative d′ of −1.02 for identifying true negatives (i.e., a marked tendency to succeed by over-accepting strings; true positive d′ = 2.75). The female model included a positive weight on phoneme.f, with negative weights on phoneme.g, phoneme.r, and phoneme.k. It classified only 6.9% of seen strings with high confidence, with a hit rate of 77% achieved at the cost of a d′ of −0.66 for identifying true positives (i.e., a tendency to over-reject strings; true negative d′ = 2.45).

The high-confidence letter models show the same pattern of over-accepting strings as *masculine* and over-rejecting them as *feminine*, resulting in the same negative d′ values as for the phoneme models, though they classified many more seen strings (about 12%). The letter models introduce several bigram predictors: in the *masculine* model, strongly negatively-weighted

[8] Apart from abbreviations and proper names, the few exceptions are limited to the words *diarrhea, errs, myrrh*, and several others containing a double *r*; *dysrhythmia* and *unrhymed*; and *tahrs*, the plural name of an Asian wild goat.

Table 9
Ten strings predicted to be highest and lowest in probability of belonging to the categories of *masculine* and *feminine*, from the 7996 strings used in the experiment.

Category	Masculine	p(Masculine)	Feminine	p(Feminine)	Masculine − Feminine	p(Masculine) − p(Feminine)
High	impicic	0.69	alyel	0.84	cruckwic	0.54
High	abonic	0.66	chalial	0.84	conctic	0.51
High	adiccon	0.66	heonia	0.81	agoxic	0.47
High	agoxic	0.66	lotial	0.81	forghic	0.45
High	alicorn	0.66	faletal	0.81	galmmic	0.45
High	allic	0.66	latolen	0.81	garphic	0.45
High	altice	0.66	bilial	0.8	gricker	0.45
High	altric	0.66	balial	0.8	doict	0.45
High	arwalic	0.66	blamial	0.8	cougzer	0.43
High	blosmic	0.66	eldial	0.8	copphic	0.42
Low	flobley	0.14	duckerk	0.13	flintay	−0.59
Low	bolfley	0.14	quask	0.13	fleipty	−0.59
Low	kiflis	0.13	corquass	0.13	flimeon	−0.6
Low	flissil	0.13	grocanx	0.13	flenia	−0.6
Low	flistry	0.1	grug	0.12	flemia	−0.6
Low	flintay	0.1	sogung	0.12	flaria	−0.6
Low	flindry	0.1	gaug	0.12	flamiant	−0.6
Low	fleipty	0.1	auggage	0.12	eaflion	−0.6
Low	flaityl	0.1	cruckwic	0.12	flaityl	−0.62
LOW	flunfle	0.09	cougzer	0.11	flissil	−0.63

Bigram.fl and strongly positively-weighted Bigram.ic, and, in the *feminine* model, strongly negatively weighted Bigram.cr and Bigram.ar. The models also include several single letters that are consistent with the phoneme model: letter.i is weighted negatively in the masculine model, and letter.k. and letter.g are weighted negatively in the feminine model.

The hybrid model for *masculine* was identical to the letter model, with its true positive rate of 63% achieved at the expense of a negative d′ (−0.97) for false positives. The high-confidence hybrid model for *feminine* classified 11% of seen stimuli, with an over-all hit rate of 70% (True positive rate = 0.34, d′ = 1.12; True negative rate = 0.36, d′ = 0.35). It used three positively weighted phonological features (CLOSE.MID, FRONT, and LATERAL.APPROXIMANT) and three negatively weighted predictors (VOICED, letter.g, and phoneme.k).

The ten most and least masculine or feminine words by this composite model are shown in Table 9.

Discussion. These models suggest that it is possible to predict strings likely to be judged *feminine* (although only a small number of strings, about 10%, can be classified with high confidence), but more difficult to accurately predict strings likely to judged *masculine*. As shown in Table 9, the most confident model judgments in female classification (p(*female*) = 0.84) are higher than the most confident model judgments in masculine classification (p(*male*) = 0.69).

The models provide some support for the idea that markers of *masculinity/femininity* are also markers of the categories *large/small* and *sharp/round* (cf. Sidhu & Pexman, 2015). The letter *k* (the fifth least common letter in English) that is negatively associated with *feminine* also had strong positive weightings in the model for the category *sharp*. The letter *g* (associated here with the category *masculine*) was also a strongly-weighted cue to membership in the category *large*. The correlations between the predicted values of the high-confidence models for each dimension support this similarity. Across all 7996 strings, the correlation between predicted *masculine/feminine* (i.e., *masculine* – *feminine*) and predicted *large/small* is 0.25 ($p < 2E-16$). The correlation between predicted *masculine/feminine* and predicted *sharp/round* is also reliable, but much lower at $r = 0.14$ ($p < 2E-16$). Together in a regression equation, estimated *large/small* and *sharp/round* account for about 12% of the variance in the *masculine/feminine* estimates ($r^2 = 0.123$, $F(2, 7993) = 562.9$; $p < 2E-16$), with the beta weight on *large/small* ($\beta = 0.32$) nearly twice that on *sharp/round* ($\beta = 0.17$).

It is not clear why femininity is more clearly symbolized than masculinity. A speculative theory is that the concept *female* is more multifaceted than the concept *male*. For instance, examining free association norms for either word reveals 12 associations for female, but only five for male (Nelson, McEvoy, & Schreiber, 1998). A more multifaceted concept might invite a more diverse set of sound symbolic associations.

General discussion: Commonly-studied categories

Using a method that does not depend on having humans make contrastive judgments, we find strong evidence of sound symbolism for three commonly-studied sound symbolism dimensions, with high-confidence hit rates equal to or greater than 70% for the categories large, sharp, round, and feminine, and above 60% for the categories small and masculine ($p < 2E-16$ in all cases, by exact binomial probability). As well as varying in their strength, these hit rates varied in where that strength came from (e.g., the hit rate for *small* is achieved mainly by a very high true positive rate of 61%, coupled with a true negative rate of just 3%, whereas the high feminine hit rate is achieved by a balance of 34% true positives and 36% true negatives). There were also large differences in the proportions of stimuli that could be classified with high confidence (defined as an absolute difference in predicted membership in either pole > 0.30). Only 15% words could be classified with high accuracy as *small*, compared to 34% of words that could be classified with high confidence as *round*.

As we noted in the introduction, it is difficult to compare the categories when they differ on so many dimensions, but it does not seem unreasonable to say that the data suggest that the categories of *sharp* and *round* are the most solidly sound symbolic, by almost any measure one might propose (e.g., confident hit rate; relative ratio of true positive and true negative classification; average d′; percent of stimuli classified with confidence multiplied by hit rate).

The hybrid models had access to features, phonemes/biphones, and letters/bigrams (though no biphone appeared in any model). It is difficult to separate these different predictor classes since they are, as different representations of the same thing, often highly or even perfectly correlated. However, there may be differences in the extent to which the high confidence models used predictors from one class or another. All models mixed predictors of all three classes (see Fig. 4). The models for *sharp/round* were relatively 'letter-heavy', with the two hybrid models using 88% (15/17)

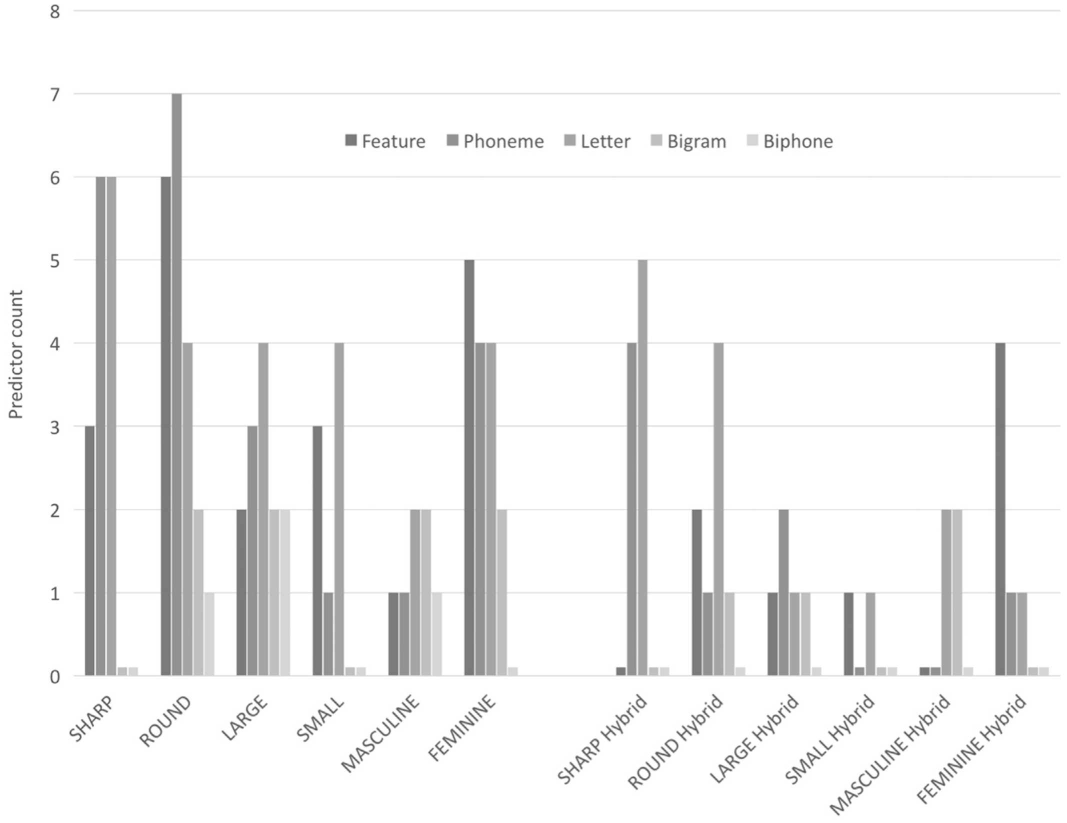

Fig. 4. Number of features, phonemes, letters, and bigrams used in modelling different dimensions. The left side shows the sum across the phoneme, letter, and bigram models. The right side shows the sum in the hybrid models only. Values of 0 are set slightly above 0 to make the categories easier to track.

unique letter/phoneme/bigram predictors, as compared to 60% (6/10) for the hybrid models for *masculine/feminine*. In contrast, the *masculine/feminine* hybrid models used 40% (4/10) unique features while the *sharp/round* models used just 11.7% (2/17) unique features. With such small numbers, it is difficult to know if these frequency differences reflect differences in the nature of the sound symbolism in the two categories. The fact that phonological features, phonemes, and letters all contribute to all sound symbolism effects complicates our understanding of the phenomena, suggesting as it does that sound symbolism may have many sources.

Semantic categories

In their discussion of the structure of semantics, Hollis and Westbury (2016) presented evidence identifying concreteness and (in keeping with earlier work by Osgood, Suci, & Tannenbaum, 1957) valence as candidate organizing dimensions.[9] There is some evidence suggesting that formal sound symbolism cues are correlated with each of these dimensions. A few studies have presented evidence of a systematic relationship between string length and concreteness (Reilly, Hung, & Westbury, 2017; Reilly, Westbury, Kean, & Peele, 2012). There is also some evidence of voiceless stops being rated as less pleasant than sonorants (Roblee & Washburn, 1912). While Greenberg and Jenkins (1966) found that front vowels were rated as less pleasant than back vowels, Miron (1961) and Tarte (1982) found the opposite pattern. Nielsen and Rendall (2011) discussed a potential evolved association between high frequency sounds (i.e., plosives, fricatives, and high-front vowels) and danger/distress, which may contribute to some of these associations.

As noted above, we tested the two dimensions of *concrete/abstract* and *good/bad* (high/low valence) in a different way than the dimensions discussed so far. Instead of asking for explicit judgments on the dimensions we selected 12 nouns that were high or low on each of these two dimensions (Brysbaert et al., 2014; Warriner et al., 2013), allowing us to define anchors for each of the two categories by aggregating the approximately 12,840 human judgments made for the six nouns that fell into that anchor's category.

Dimension 5: Concrete/Abstract

The full hybrid dimension of *concrete/abstract* was defined by concatenating the judged categories *wisdom, spirituality, virtue, sadness, fraud,* and *injustice* to construct the abstract category, and by concatenating the judged categories *flower, gem, toy, wasp, bomb,* and *fungus* to construct the concrete category.

The best models of *concrete* and *abstract* are shown in Table 10. The models may be summed up in the briefest way by saying that they are very poor models. The hybrid models for both poles include true positive d′ values at or below zero (Concrete true positive d′ = 0; Abstract true positive d′ = −0.07). The high confidence models classify fewer than ten stimuli, and all incorrectly.

Since all this evidence suggests that the models perform only marginally better than chance, we have not included a table of the few nonwords classified.

We defer discussion until consideration of the second semantic category, which showed very similar results.

Dimension 6: High/Low valence

The full hybrid dimension of *high/low valence* was defined by concatenating the judged categories *wisdom, spirituality, virtue, flower, gem* and *toy* to construct the high valence category, and

[9] Other dimensions were associated with concepts that are not easily suited for a sound symbolism study: *word frequency*, *agency*, and *meaning specificity*.

Table 10

Model summary for categories *concrete* and *abstract*. Predictors are ordered by decreasing magnitude of beta weight. Predictors with a positive weight are shown in bold. **CV**: K-fold cross-validated hit rate (k = 10). **Hits**: Exact observed hit rate. **TP**: True positive rate. **TN**: True negative rate. **Difference models** are limited to high confidence strings, defined as estimated to have a difference in probability of belonging to one category (*concrete* or *abstract*) – probability of belonging the other pole > 0.30. **Proportion**: Proportion of seen strings in the difference model.

Model	Concrete	Estimate	SE	t	Pr(>|t|)	Performance	Difference model	Abstract	Estimate	SE	t	Pr(>|t|)	Performance	Difference model
Features	**(Intercept)**	**0.03**	**0.03**	**0.99**	**3.20E−01**	CV: 0.52	PROPORTION: 0	(Intercept)	−0.34	0.05	−6.5	8.36E−11	CV: 0.55	PROPORTION: 0
	CLOSE.MID	**0.17**	**0.03**	**5.17**	**2.41E−07**	HITS: 0.52	HITS: 0	**CLOSE.MID**	**0.24**	**0.03**	**7.13**	**1.08E−12**	HITS: 0.55	HITS: 0
	VOICED	−0.1	0.02	−4.28	1.89E−05	TP: 0.34	TP: 0	LATERAL.APPROXIMANT	−0.23	0.04	−6.45	1.14E−10	TP: 0.18	TP: 0
						TP d′: 0.56	TP d′: N/A	**FRONT**	**0.21**	**0.03**	**8.11**	**5.51E−16**	TP d′: −0.38	TP d′: N/A
						TN: 0.18	TN: 0	**OPEN**	**0.18**	**0.04**	**4.31**	**1.68E−05**	TN: 0.37	TN: 0
						TN d′: −0.42	TN d′: N/A	PLOSIVE	−0.12	0.02	−5.68	1.40E−08	TN d′: 0.69	TN d′: N/A
								ALVEOLAR	**0.1**	**0.02**	**4.77**	**1.88E−06**		
Phonemes	(Intercept)	−0.03	0.02	−1.35	0.18	CV:	PROPORTION: 0	(Intercept)	−0.03	0.03	−0.98	0.33	CV:	PROPORTION: 0.0002
	Biphone.nt	−0.34	0.09	−3.92	8.88E−05	HITS: 0.53	HITS: 0.5	Biphone.kAX	−0.72	0.12	−6.26	4.10E−10	HITS: 0.52	HITS: 0.5
	Phoneme.AX	**0.17**	**0.03**	**5.27**	**1.43E−07**	TP: 0.23	TP: 0.5	Biphone.AXr	−0.45	0.09	−5.16	2.49E−07	TP: 0.21	TP: 0.5
						TP d′: −0.16	TP d′: N/A	Biphone.IHk	0.4	0.09	4.33	1.50E−05	TP d′: −0.2	TP d′: N/A
						TN: 0.3	TN: 0	Biphone.AXs	0.36	0.09	4.1	4.18E−05	TN: 0.31	TN: 0
						TN d′: 0.33	TN d′: N/A	Phoneme.g	−0.31	0.05	−5.93	3.03E−09	TN d′: 0.31	TN d′: N/A
								Phoneme.AX	0.28	0.04	7.27	3.81E−13		
								Phoneme.f	−0.25	0.05	−5.21	1.93E−07		
								Phoneme.b	−0.21	0.04	−4.7	2.68E−06		
								Phoneme.l	−0.14	0.03	−4.23	2.38E−05		
Letters	(Intercept)	−0.01	0.03	−0.49	0.63	CV: 0.52	PROPORTION: 0.0009	(Intercept)	−0.12	0.03	−3.84	0.000122	CV: 0.56	PROPORTION: 0.0008
	Bigram.um	**0.58**	**0.11**	**5.4**	**6.98E−08**	HITS: 0.53	HITS: 0.36	Bigram.fl	−0.66	0.11	−5.81	6.34E−09	HITS: 0.56	HITS: 0.8
	Bigram.us	**0.43**	**0.09**	**4.93**	**8.30E−07**	TP: 0.25	TP: 0.36	Letter.k	−0.46	0.06	−7.82	5.72E−15	TP: 0.25	TP: 0
	Letter.u	−0.23	0.05	−4.96	7.20E−07	TP d′: 0	TP d′: N/A	Letter.z	−0.41	0.08	−4.87	1.15E−06	TP d′: 0.05	TP d′: N/A
	Letter.a	**0.12**	**0.03**	**4.23**	**2.39E−05**	TN: 0.27	TN: 0	Letter.w	−0.4	0.1	−3.93	8.60E−05	TN: 0.31	TN: 0.8
						TN d′: 0.17	TN d′: N/A	**Bigram.us**	**0.36**	**0.08**	**4.73**	**2.26E−06**	TN d′: 0.34	TN d′: N/A
								Letter.b	−0.24	0.04	−6.01	1.95E−09		
								Letter.g	−0.24	0.04	−5.58	2.46E−08		
								Letter.i	**0.24**	**0.03**	**7.65**	**2.08E−14**		
								Letter.a	**0.19**	**0.03**	**6.82**	**9.75E−12**		
Composite	**(Intercept)**	**−0.01**	**0.03**	**−0.49**	6.30E−01	CV: 0.52	PROPORTION: 0.0001	(Intercept)	−0.1	0.03	−2.91	0	CV:	PROPORTION: 0.0001
	Bigram.um	**0.58**	**0.11**	**5.4**	**6.98E−08**	HITS: 0.53	HITS: 0	Biphone.AXs	0.53	0.08	6.28	3.61E−10	HITS: 0.57	HITS: 0
	Bigram.us	**0.43**	**0.09**	**4.93**	**8.30E−07**	TP: 0.25	TP: 0	Letter.k	−0.46	0.06	−7.83	5.36E−15	TP: 0.23	TP: 0
	Letter.u	−0.23	0.05	−4.96	7.20E−07	TP d′: 0	TP d′: N/A	Letter.w	−0.42	0.1	−4.1	4.21E−05	TP d′: −0.07	TP d′: N/A
	Letter.a	**0.12**	**0.03**	**4.23**	**2.39E−05**	TN: 0.27	TN: 0	Letter.z	−0.42	0.09	−4.96	7.08E−07	TN: 0.34	TN: 0
						TN d′: 0.17	TN d′: N/A	Phoneme.g	−0.35	0.05	−6.65	3.06E−11	TN d′: 0.48	TN d′: N/A
								Phoneme.f	−0.29	0.05	−6.01	1.89E−09		
								Letter.b	−0.26	0.04	−6.39	1.73E−10		
								Letter.i	0.25	0.03	7.88	3.51E−15		
								Letter.a	0.19	0.03	6.68	2.58E−11		

Table 11
Model summary for categories *high valence* and *low valence*. Predictors are ordered by decreasing magnitude of beta weight. Predictors with a positive weight are shown in bold. **CV**: K-fold cross-validated hit rate (k = 10). **Hits**: Exact observed hit rate. **TP**: True positive rate. **TN**: True negative rate. **Difference models** are limited to high confidence strings, defined as estimated to have a difference in probability of belonging to one category (*high valence* or *low valence*) – probability of belonging the other pole > 0.30. **Proportion**: Proportion of seen strings in the difference model.

Model	High valence	Estimate	SE	t	Pr(>|t|)	Performance	Difference model	Low valence	Estimate	SE	t	Pr(>|t|)	Performance	Difference model
Features	(Intercept)	−0.23	0.07	−3.32	9.08E−04	CV: 0.564	PROPORTION: 0	[No model]					[No model]	[No model]
	FRONT	0.34	0.03	11.12	<2E−16	HITS: 0.57	HITS: 0							
	CLOSE.MID	0.33	0.04	9.16	<2E−16	TP: 0.26	TP: 0							
	LABIODENTAL	−0.3	0.04	−6.85	8.01E−12	TP d′: 0.16	TP d′: N/A							
	BACK	0.21	0.04	5.41	6.29E−08	TN: 0.31	TN: 0							
	VELAR	−0.16	0.03	−5.13	2.97E−07	TN d′: 0.27	TN d′: N/A							
	PLOSIVE	−0.12	0.02	−5.06	4.18E−07									
	VOICED	−0.09	0.02	−5.1	3.37E−07									
Phonemes	(Intercept)	−0.21	0.04	−5.66	1.55E−08	CV: 0.56	PROPORTION: 0	(Intercept)	−0.02	0.02	−1.06	0.29	CV: 0.529	PROPORTION: 0
	Biphone.AXr	−0.43	0.09	−4.95	7.55E−07	HITS: 0.56	HITS: 0	**Biphone.AXn**	**0.21**	**0.05**	**3.87**	**0.00011**	HITS: 0.53	HITS: 0
	Phoneme.EY	**0.39**	**0.08**	**4.68**	**2.89E−06**	TP: 0.19	TP: 0	**Phoneme.s**	**0.19**	**0.04**	**4.99**	**6.23E−07**	TP: 0.26	TP: 0
	Phoneme.g	−0.37	0.05	−6.89	5.82E−12	TP d′: −0.31	TP d′: N/A	**Phoneme.m**	**0.18**	**0.04**	**4.16**	**3.19E−05**	TP d′: 0.01	TP d′: N/A
	Phoneme.AX	**0.36**	**0.04**	**9.28**	**<2E−16**	TN: 0.37	TN: 0						TN: 0.27	TN: 0
	Phoneme.f	−0.34	0.05	−6.88	6.48E−12	TN d′: 0.69	TN d′: N/A						TN d′: 0.17	TN d′: N/A
	biphone.AXl	−0.27	0.06	−4.32	1.58E−05									
	Phoneme.IY	**0.23**	**0.04**	**5.22**	**1.79E−07**									
	Phoneme.b	−0.23	0.04	−5.27	1.41E−07									
	Phoneme.AE	**0.21**	**0.04**	**4.8**	**1.57E−06**									
	Phoneme.IH	**0.21**	**0.04**	**5.55**	**2.95E−08**									
	Phoneme.k	−0.19	0.04	−5.37	7.98E−08									
Letters	(Intercept)	−0.17	0.03	−4.92	8.97E−07	CV: 0.564	PROPORTION: 0.0003	**(Intercept)**	**0.04**	**0.02**	**1.8**	**0.07**	CV: 0.529	PROPORTION: 0.0009
	Letter.k	−0.4	0.06	−6.74	1.70E−11	HITS: 0.56	HITS: 0.75	**Bigram.si**	**0.44**	**0.11**	**3.86**	**0.000115**	HITS: 0.53	HITS: 0.33
	Letter.g	−0.34	0.04	−7.67	1.82E−14	TP: 0.2	TP: 0	Bigram.ee	−0.43	0.1	−4.25	2.15E−05	TP: 0.49	TP: 0.33
	Letter.f	−0.3	0.04	−6.73	1.81E−11	TP d′: −0.23	TP d′: N/A	Bigram.ch	−0.37	0.09	−4.09	4.36E−05	TP d′: 1.83	TP d′: N/A
	Bigram.us	**0.3**	**0.08**	**3.99**	**6.60E−05**	TN: 0.36	TN: 0.75	**Bigram.st**	**0.31**	**0.08**	**3.86**	**0.000115**	TN: 0.04	TN: 0
	Letter.a	**0.25**	**0.03**	**9.04**	**<2E−16**	TN d′: 0.61	TN d′: N/A	**Letter.c**	**0.15**	**0.04**	**4.16**	**3.21E−05**	TN d′: −1.59	TN d′: N/A
	Letter.i	**0.24**	**0.03**	**7.61**	**2.87E−14**									
	Letter.b	−0.21	0.04	−5.33	1.01E−07									
	Letter.d	−0.17	0.04	−3.84	0.000125									
Composite	(Intercept)	−0.12	0.05	−2.48	0.01	CV: 0.57	PROPORTION: 0.0007	(Intercept)	0	0.02	0.07	0.95	CV: 0.523	PROPORTION: 0.0009
	Biphone.AXr	−0.43	0.09	−4.99	6.14E−07	HITS: 0.57	HITS: 0.67	**Bigram.si**	**0.46**	**0.11**	**4**	**6.31E−05**	HITS: 0.53	HITS: 0.5
	Phoneme.f	−0.39	0.05	−7.71	1.37E−14	TP: 0.19	TP: 0	Bigram.ee	−0.42	0.1	−4.11	4.04E−05	TP: 0.49	TP: 0.5
	Phoneme.AX	**0.36**	**0.04**	**9.42**	**<2E−16**	TP d′: −0.28	TP d′: N/A	Bigram.ch	−0.37	0.09	−4.02	5.86E−05	TP d′: 1.83	TP d′: N/A
	Letter.k	−0.32	0.06	−5.37	7.83E−08	TN: 0.38	TN: 0.67	**Bigram.st**	**0.32**	**0.08**	**4.02**	**5.92E−05**	TN: 0.04	TN: 0
	Biphone.AXl	−0.29	0.06	−4.55	5.55E−06	TN d′: 0.72	TN d′: N/A	**Phoneme.m**	**0.18**	**0.04**	**4**	**6.44E−05**	TN d′: −1.59	TN d′: N/A
	Letter.g	−0.28	0.04	−6.3	3.13E−10			**Letter.c**	**0.16**	**0.04**	**4.4**	**1.08E−05**		
	FRONT	**0.23**	**0.03**	**9.05**	**<2E−16**									
	PLOSIVE	−0.16	0.02	−7.07	1.64E−12									

Table 12
Hybrid model performance for individual categories making up the imageability and valence dimensions.

Category	Hit rate	*p*	TP	TP d′	TN	TN d′
spirituality	0.62	<2E−16	0.23	0.28	0.40	1.18
flower	0.62	<2E−16	0.40	1.15	0.22	−0.03
toy	0.61	<2E−16	0.00	N/A	0.61	N/A
wisdom	0.59	<2E−16	0.25	0.11	0.34	0.88
virtue	0.58	7.09E−13	0.23	0.07	0.35	0.64
sadness	0.55	8.70E−07	0.46	2.73	0.09	−1.85
gem	0.55	5.63E−06	0.23	−0.16	0.32	0.62
fraud	0.55	8.36E−06	0.15	−1.15	0.40	1.89
wasp	0.54	1.80E−05	0.47	3.03	0.07	−2.10
bomb	0.51	0.14	0.00	N/A	0.51	5.56
fungus	No model	No model	N/A	N/A	N/A	N/A
injustice	No model	No model	N/A	N/A	N/A	N/A

Table 13
Hybrid model for *flower.*

	Estimate	SE	t	*p*
(Intercept)	−0.03	0.1	−0.29	0.77
Phoneme.AX	0.57	0.09	6.62	4.55E−11
VELAR	−0.42	0.07	−5.91	4.08E−09
FRONT	0.37	0.06	5.99	2.40E−09
Phoneme.t	−0.35	0.09	−3.95	8.22E−05

Table 14
Ten strings predicted to be highest and lowest in probability of belonging to the category of *flower*, from the 7996 strings used in the experiment.

Category	String	p(FLOWER)
High	asanism	0.87
High	eponism	0.87
High	heonia	0.87
High	amisism	0.84
High	etonism	0.82
High	adelous	0.82
High	aerble	0.82
High	aerson	0.82
High	airamus	0.82
High	aromal	0.82
Low	conctic	0.22
Low	corquass	0.22
Low	counk	0.22
Low	glonk	0.22
Low	glonx	0.22
Low	goonx	0.22
Low	quask	0.22
Low	woonc	0.22
Low	woonk	0.22
Low	cruckwic	0.21

Table 15
All 31 hybrid model predictors, sorted in descending order of the average absolute β weights assigned to them.

PREDICTOR	AVERAGE |β|	MODEL[S]
Letter.x	1.81	SHARP
Bigram.fl	1.25	MALE
Bigram.oo	1.06	LARGE
Bigram.am	1.04	ROUND
Bigram.ic	0.9	MALE
Phoneme.g	0.88	LARGE
Phoneme.OW	0.77	SHARP
Phoneme.UW	0.77	SHARP
Letter.k	0.76	SHARP
Letter.y	0.69	MALE
Letter.h	0.66	SMALL
Phoneme.k	0.59	FEMALE
Letter.o	0.58	ROUND
Letter.c	0.58	SHARP
Phoneme.b	0.57	LARGE
Phoneme.AX	0.57	FLOWER
Letter.m	0.52	SHARP
Letter.g	0.5	FEMALE
CLOSE.MID	0.48	FEMALE
LABIODENTAL	0.48	ROUND
Letter.u	0.47	ROUND
Phoneme.AA	0.45	SHARP
Letter.t	0.445	ROUND/SHARP
FRONT	0.42	FEMALE/FLOWER
VELAR	0.42	FLOWER
k	0.41	ROUND
Letter.i	0.39	MALE
Phoneme.t	0.35	FLOWER
LATERAL.APPROXIMANT	0.345	FEMALE/ROUND
BACK	0.33	LARGE
VOICED	0.28	FEMALE/SMALL

by concatenating the judged categories *wasp, bomb, fungus, sadness, fraud,* and *injustice* to construct the low valence category.

The full models are shown in Table 11. The hybrid model for high valence achieved a negative true positive d′ of −0.28. The low valence model is strongly biased to accepting strings, with true positive rate of 0.49 (d′ = 1.83) at the expense of the true negative rate of 0.04 (d′ = −1.59). It classifies only eight stimuli with high confidence with a 50% hit rate. As with the model above, we have not included a table of classified nonwords since the model is so poor.

Discussion. These two attempts to document sound symbolism indirectly were failures (but see Louwerse and Qu (2017), for a successful, more direct approach to sound symbolism in valence). To understand our results, we developed models on each of the component categories (which are, recall, the categories on which the participants actually judged membership). Table 12 shows hybrid model hit rates for models of string acceptance that were developed on each of the 12 component categories. Nine of the twelve models (all but *bomb, fungus,* and *injustice*) showed statistically reliable hit-rates, ranging from a 62% for *spirituality* to 54% for *wasp*. However, of these, all but one are trivially high, either because they are strongly biased to acceptance (*wasp* and *sadness*, with high true positive d′ scores but strongly negative true negative d′ scores) or to rejection (*toy* and *fraud*, with high true negative rates but very poor true positive rates), or because the true positive d′ values are very low or negative (*gem, spirituality, virtue, wisdom*).

The sole exception is the category *flower*, with a total hit rate of 0.62, which has a good true positive hit rate (0.40) and good true positive d′ (1.15), though a near-zero d′ for true negatives (−0.03; True negative hit rate: 0.22). The model for flower is shown in Table 13, with the most and least likely strings picked out by the model shown in Table 14. The list includes several morphologically-implausible words that would probably be rejected as flower names by human judges because of the suffix

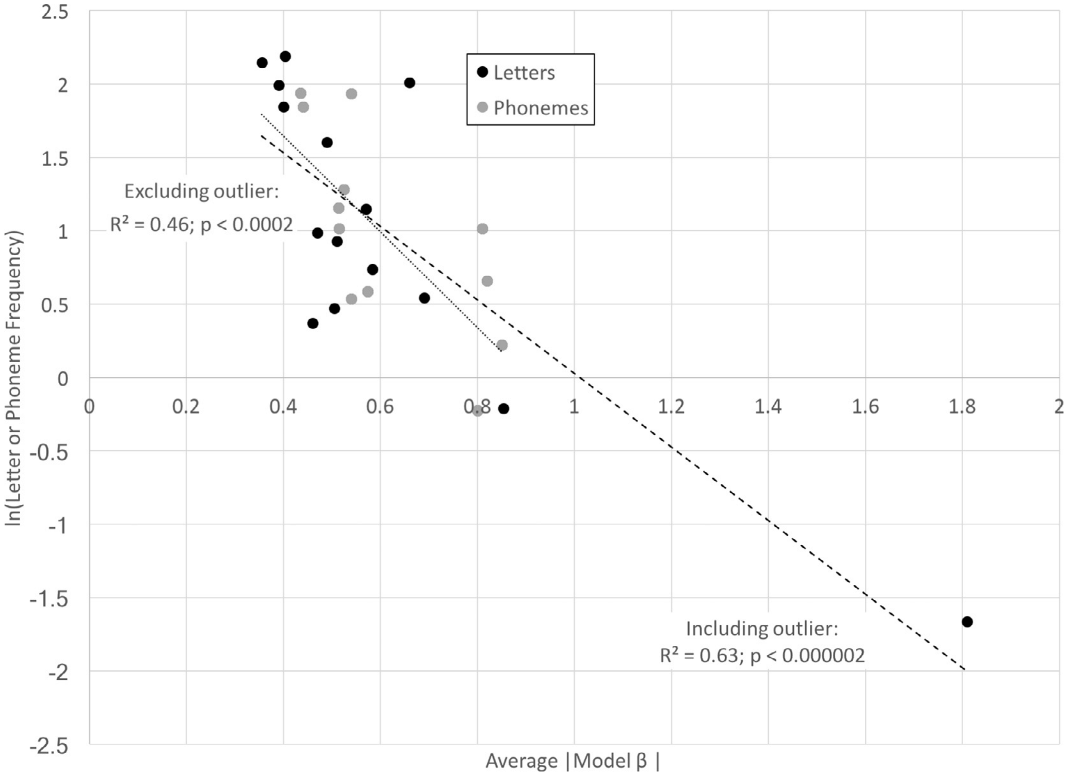

Fig. 5. Correlation between logged letter and phoneme frequencies and the average absolute beta weights assigned to those predictors in any hybrid model in which they appeared. The outlier is letter.x.

ism, which is generally used in English to form nouns referring to actions or completed acts (*bilingualism, ventriloquism, witticism*) or to conduct (*Buddhism, egoism, sadism*). The model has two positively-weighted predictors that both favor small vowels: phoneme.AX [/ə/] (β = 0.57) and FRONT (β = 0.37). It also has two negatively weighted predictors: VELAR (β = −0.42) and phoneme.t (β = −0.35). These predictors may be explained by noting that many of the (negatively-weighted) VELAR consonants *g* and *k* have associations with *large*, (negatively-weighted) phoneme.t has a strong association with *sharp*, and FRONT is associated with *small*. For converging evidence about the validity of the model for *flower*, we can examine how model predictions in this category related to predictions for categories considered above that are semantically related to flowers. Concretely, we can expect that strings suggested by the model as suitable for *flower* should be more closely related to the categories *small*, *feminine*, and *round* than to the categories anchoring the opposite poles of *large*, *masculine*, *sharp* (although of course a small subset of flowers are strongly associated with sharpness due to having thorns). Across all 7996 strings, the correlation between predicted *masculine/feminine* and predicted *flower* was $r = -0.70$ ($p < 2E-16$); between flower and *sharp/round* was −0.25 ($p < 2E-16$); and between flower and *large/small* was −0.04 ($p < 0.0005$). All of these relationships are in the direction one would expect given the semantic relationships of the categories with the word *flower*. A regression model predicting the values estimates for *flower* from the three others was highly reliable ($F(3, 7992) = 2911$, $p < 2E-16$), and achieved an r^2 value of 0.52, with all three predictors entering in reliably: *masculine/feminine* β = −0.44, *sharp/round* β = −0.05, and *large/small* β = −0.05. See also Fig. 3, which illustrates the very strong positive correlation of $r = 0.77$ ($p < 2E-16$) between predicted *femininity* and predicted *flower*. This suggests again that some sound symbolic effects are mediated by other phoneme-meaning associations (see French, 1977).

General discussion: Semantic categories

In the introduction to this paper, we posed several unanswered questions about sound symbolism that we hoped to be able to address.

The first question was: *Is there any systematicity to which linguistics features can act as sound symbols?* We made several explicit predictions about the nature of the cues that could be sound symbolic. One was that sound symbolism cues should be useful as discriminative signs in inverse proportion to their frequency. We were able to test this prediction because we have the β weights of every feature, phoneme, and letter that entered into a model. Table 15 shows a list of all 31 predictors that appeared (only a few more than once) in any of the successful hybrid models, sorted in order of decreasing average absolute beta weight.[10] It is immediately obvious from inspection that less common features were assigned higher weights, as predicted. The five most highly weighted predictors are the low frequency letter letter.x (third least common letter) and the only four bigrams (necessarily low frequency compared to letters) that appeared in any model. The five least weighted features include (as predicted) ubiquitous phonological features (most notably, BACK and VOICED) and letter.t (the second most common letter, after e).

The logged frequencies of the letters (computed from approximately 4.5 billion characters of English text by Lyons, n.d.) and phonemes[11] (from Blumeyer, 2012) are graphed against their average absolute model weights in Fig. 5. The values have a very strong negative correlation ($r = -0.83$, $p = .000006$), i.e., less frequent char-

[10] This measure is slightly misleading in some cases, since a few models contain oppositely-signed weights on phonemes and their orthographic representation, which raises their functional weight in that particular model. However, it is a reasonable rough estimate of the discriminative value of each cue.

[11] Plus the feature LATERAL.APPROXIMANT, since it encodes only one English phoneme, /l/.

acters are not only more strongly weighted as predictors; they are so weighted in direct inverse proportion to their logged frequency. As we suggested in the introduction, this is predictable from an information theoretic point of view, since infrequent characters are, in virtue of their infrequency, better discriminative cues.

The notion that low frequency cues carry more sound symbolic weight also has some precedence in the research on ideophones (for a review see Dingemanse, 2012). These are iconic words present in many languages (though uncommon in Indo-European languages) that depict various sensory events. This is often accomplished via sound symbolic links between the form of the ideophone and the sensory event being depicted. For instance, the Japanese word *pika* means a flash of light–note that it contains the vowel /i/ which is associated with brightness (Newman, 1933). Dingemanse (2012) notes that ideophones are often marked by an unusual form. While they do not contain unusual phonemes, these phonemes are often arranged in unusual ways, leading to skewed phonotactic distributions and greater variability in syllable structure. Lockwood and Tuomainen (2015) theorize that their markedness triggers a "sensory integration process of sound and the sensory information which the ideophone activates by association" (p. 7). We might theorize that this is also involved in the greater sound symbolic effect among low frequency items observed here. When a given component (i.e., feature, phoneme, grapheme) is marked due to its low frequency, it may be processed in such a way that allows for its sound symbolic associations to affect the interpretation of its nonword. Lockwood and Tuomainen (2015) observed a larger P2 for ideophones, which they interpreted as reflecting this process. Future research might examine if there is any evidence of a similar effect for the components identified here.

We also predicted that, *pari passu*, letters with ambiguous pronunciation should be weighted lower than similar letters with unambiguous pronunciation. We can test this in addressing the question we raised earlier: Why does phoneme.b act more strongly as a sound symbolic cue than other voiced plosives? Phoneme.b and letter.g each appeared, with mid-range weights, in a single hybrid model, but the third voiced plosive *d* was not seen in any model and is not commonly associated with sound symbolism. The non-symbolic phoneme.d is far more common (accounting for 4.21% of all phonemes) than either phoneme.g (0.8%) or phoneme.b (1.8%). In virtue of being more commonly experienced (i.e., of appearing in a wide range of words, and therefore being less likely to be a cue to any particular meaning) it is predictable that it should have lower value as a discriminative cue than other voiced plosives, as it does.

We would also predict that the symbolic value of letter.g should be further diluted by its greater phonological ambiguity, compared to letter.b. Although both letter.b and letter.g have more than one pronunciation (e.g., *bog* vs. *lamb*; *game* vs. *age*), a silent letter.b pronunciation is far less common than letter.g as /dʒ/. In an English dictionary of 111,626 words (Shaoul & Westbury, 2006), there are 514 lemmas that end with 'ge', suggesting that with their morphological families there are hundreds of words in which 'g' is pronounced /dʒ/. In contrast, there are just 37 lemmas that end with a silent 'mb', so (taking into account their morphological families) at most a few dozen words with a silent letter.b. There are over a thousand words containing 'mb' in which the letter.b is not silent: i.e., *number, November, amber, ambition, remember*. These distributions suggest that letter.b, with its more consistent mapping to phonology, might be expected to be a more reliable discriminator of sound symbolic meaning than letter.g. However, the evidence that this is true is not strong. In the models in which they appear, phoneme.b has a slightly higher absolute weight (0.57) than letter.g (0.5), a difference of 0.24z in terms of the normalized absolute weights of letters and phonemes.

We also suggested that vowels-as-letters should be weak cues due to the facts that they are common and have multiple pronunciations. The four vowels-as-letters that appear in any of the models (letter.a, letter.i, letter.o, and letter.u) have an average absolute beta weight of 0.47, as compared to an average absolute beta weight of 0.69 for the nine consonants-as-letters.[12] The consonant average includes the predictably low weight for letter.t, which is more common in English than every vowel except letter.e, and as a result has a low absolute beta weight of 0.4. If we ignore it as an outlying consonant, the average absolute beta weight for consonants is 0.72. This vowel-consonant weight difference is a large difference of 0.87z in terms of the normalized absolute weights on all letters and phonemes.

Although this strong relationship between the frequency of a predictor and its strength as a sound symbolic cue does not say anything about the semantics of sound symbolism (i.e., about what leads to particular cues being associated with *particular* semantic dimensions), it does place strong constraints on those semantics. There are only a small number of low probability cues that can act as strong sound symbols. Since they cannot be sound symbolic for contradictory categories (which would make them useless as discriminatory cues), they must be distributed in a way that restricts contradictory symbolism. The consequence of having a small number of strong cues that must be distributed quasi-independently across categories is that there cannot be a large number of sound symbolism categories with strong cues. Sound symbolism with weak (i.e., frequent) cues can exist, of course, but it too must be very limited, for two reasons: because the number of weak cues that can be combined is capped by word length restrictions and because combining numerous weak cues would rapidly become non-discriminatory. If there are many weakly-weighted cues defining a category, one of two things must be true. One is that the category will rely on only a few of these weakly-weighted cues and thus by definition be defined only with low certainty. The other is that the category will rely on combining a large number of cues, and (since the cues are by assumption common) too many strings will be able to meet the criteria so the category will be very large and diverse. In either case, the category will not be a clearly discernible category.

There are a couple of corollaries to these restrictions. The first is that, since a small number of low-frequency cues are available to symbolize a number of potentially contradictory semantic categories, the more low-frequency cues a string contains, the higher the probability that its symbolic interpretation will be indeterminate, since there is a higher probability of the string's cues sending contradictory signals. In order to test this, we split the alphabet in half by letter frequency and used binomial regression to predict the responses to all 42,778 decisions, without regard to semantic category, using a count of the high and low frequency letters as predictors. Only the count of low frequency letters entered the model, with a negative $\beta = -0.13$ ($p < 2\text{E}-16$). As predicted, a greater number of low frequency (only) letters in a string is associated with a lower acceptance rate for that string.

A second interesting corollary of these restrictions is that they explicitly suggest where there may yet be untapped potential to discover sound symbolism: among the few rare cues that are yet 'spoken for'. Although some of the rare letters and phonemes that are not strongly associated with any known sound symbolism category may not be associated precisely because they are too rare to be useful cues, some may be able to play the role. The remaining rare cues are letters *z* and *j* and their associated phonemes /ʒ/ and /dʒ/, as well as unvoiced th /θ/, *ch* (/tʃ/), and perhaps /aʊ/ (*ow* as in *cow)* and /ʊ/ (*oo* as in *foot*), though both these vowel pho-

[12] We ignored letter.y due to its ambiguous role.

nemes have characteristics that would allow them to be sound symbolic in known categories, notably *large* and *small*, respectively.

The second question we posed at the beginning of this paper was: *Are sound symbolic effects limited to the commonly-studied dimensions, or are they more general?* We found only weak evidence for generalization outside of the best known dimensions. Neither of the hybrid categories, *concreteness* or *valence*, were successfully modeled. Of their twelve component categories, only *flower* was successfully modeled. This seems to have been largely because the category is strongly associated with several other well-symbolized categories, with the estimates most notably correlated at $r = -0.70$ with the estimates for *masculine/feminin*e (and correlating at $r = 0.77$ for estimates of *feminine* alone), as well as correlating reliably and negatively with *large/small* and *sharp/round* ($p < 2E-16$ in all cases). These twelve component categories may be too specific to result in sound symbolic associations. Monaghan, Mattock, and Walker (2012) have provided some evidence that sound symbolism operates at broad categorical levels (e.g., an association with *roundness* in general, as opposed to specific round entities). Thus, while sound symbolism was not found to generalize to specific target dimensions here, there are a number of other broad categorical associations (e.g., *fast/slow*, *bright/dull*) that have been demonstrated in the literature. It remains to be seen if such associations can be observed using this large-scale approach.

It is possible, however, that a small number of basic associations (potentially just those involving the most strongly-symbolized dimensions we have considered here) underlie the majority of sound symbolic associations (e.g., French, 1977). For instance, the association between certain vowels and *largeness* might facilitate an association between those vowels and other dimensions related to largeness (e.g., *thickness* or *slowness*). We found some evidence of this here. Fig. 3 speaks to the semantic question (*why do the sound symbols that exist play their role?*) by suggesting more specifically that the category *large* may lend its sound symbolic power to several categories. As we have noted above, estimates for *large* are negatively correlated with estimates for *small, sharp, flower*, and *feminine*, and reliably positively correlated with *round*. The category *large* is also the most predictable of the categories we used in our experiments (Fig. 1), although entirely on the basis of true negatives (that dovetail with strong true positive performance in the category *small*). As we have also noted above, we are tempted to argue that this *large/small* dimension is the only dimension definable using phonological features alone. It has long been noted by others, going back to Jespersen (1925), that the back vowels and voicing that are associated with *large* are both sounds with a natural counterpart in non-linguistic domains, since big animals (and, more generally, big things when struck or used as horns) make louder, lower sounds (see also Fitch, 1997; Ohala, 1983; Ohala, 1984). Its definability with phonological features may reflect that *large/small* is a true 'natural' category whose existence helps delimit the several other categories with which it is correlated.

The third question we posed was: *Are the predictors of one pole of the dimensions the same (with reversed sign) as the predictors of the other pole, or are the poles separable?* Our results suggest that the answer is that the poles are largely separable. Perhaps the best example is the two poles of the masculine/feminine dimension, which show a large difference in predictability (masculine high confidence hit rate of 63% versus a feminine high confidence hit rate of 70%). Moreover, in several cases our models suggested that one pole of a dimension was predicted largely in terms of being the negative of the other, implying that it is mainly one pole of the dimension that is responsible for the sound symbolism. As discussed in the last paragraph, a clear example is the dimension of large/small, which may be best described as a dimension of large or not-large, since all predictors in all the models for small were negatively weighted (although, nevertheless, these models do well at predicting true positives in the category small, due to a bias [i.e., positive intercept] towards accepting strings as small in the absence of any counter-evidence).

It is interesting to consider to what extent similar one-sided anchors might define some semantic dimensions, and whether this is because some poles are more salient because of perceptual, biological, linguistic, or other factors. The idea of opposing anchors of a dimension being differently represented is related to *markedness*. This complex construct may be composed of many quasi-independent cues (Lehrer, 1985), but is essentially the extent to which the extent to which one of a pair of antonyms is used as a default (unmarked) or not used as a default (marked). However, the direct relevance is unclear due to inconsistencies. By the wide variety of cues reviewed by Lehrer, largeness is unmarked (and was positively symbolized in our study), and smallness is marked (and was poorly symbolized in our study, mostly by its absence). On the other hand, by the same criteria maleness is unmarked and femininity is marked, but femininity was better symbolized in our data than masculinity. It may be possible to relate these differences to differences in the cues that contribute to markedness.

Notably, the distinction between the meanings of two opposing anchors of a dimension is obscured in studies that present participants with both ends of a given dimension (e.g., studies in which participants are given two nonwords to match with a round *and* a sharp visual shape; see French, 1977). A similar result was reported by D'Onofrio (2014), who found that while some features (e.g., the presence of voicing) were associated with round shapes, their opposites (e.g., the absence of voicing) weren't necessarily associated with sharp shapes.

This question of whether the poles of a sound symbolized dimension are separable can perhaps be more succinctly addressed by simply noting again the correlations between predictions for opposite poles. While estimates for all opposing poles are reliably negatively correlated, as one would expect, the magnitude of the correlations are not as high as one might expect for 'true opposites' (sharp:round: $r = -0.66$; large:small $r = -0.34$; masculine:feminine: $r = -0.34$).

Our fourth question was: *Are all form predictors in a predictable semantic dimension equal in their predictive force?* As we have discussed above, our models suggest clearly that the answer is no. We see a wide range of weights and signs in the beta weights in the models, and those weights are strongly correlated with the logged frequencies of their associated predictors (Fig. 5).

Our fifth and sixth questions concerned the nature of those form cues. The fifth question was: *Are the effects entirely phonological or might they also (or rather) be orthographic?* We have suggested that in general the phonological ambiguity of much orthography suggests that phonology is likely to be a better discriminant cue. However, we found both orthographic and phonological cues in our models, in almost equal numbers. Altogether there were 31 predictors used in the hybrid models, 14 of which were graphemes and 13 of which were phonemes. Because we arbitrarily deleted the orthographic cues when orthographic and phonological cues were highly correlated, this may slightly overestimate the weight placed on phonological cues. It would be interesting (though perhaps very difficult or impossible) to identify dimensions better predicted by one or the other and see if the sound symbolism cues dissociated as the discriminatory power of the cues suggests they would (see Cuskley, Simner, & Kirby, 2015; Kirby, 2015; Sidhu et al., 2016 for attempts at distinguishing the roles of orthography and phonology in shape sound symbolism).

The final question we posed was: *Do biphones or bigrams contribute to the effects, and if so, how strongly?* Our 31 cues included no biphones and just four bigrams. All the bigrams that did appear in a final model were very strongly weighted, as would be predicted from the fact that they are necessarily rare and therefore highly discriminative cues (Average absolute beta weight: 1.06; no weight was lower than 0.9). In keeping with the discussion above of where else we might see undiscovered sound symbolism, we suspect that bigrams may also be a fruitful avenue to explore, although many bigrams may be too rare to serve as useful discriminant cues, since cues that are hardly ever encountered are not useful cues.

Conclusion

The approach we have taken in this study suggests that we can cleave the sound symbolism problem into two parts.

One part is the semantics of the situation: i.e., *why* certain specific cues symbolize specific categories. While several theories have been proposed, there is still much work to be done adjudicating between them. In addition, the majority of work on sound symbolism has focused on specific subsets of language. We suggest that the field could benefit from the large-scale approach taken here, to test and refine existing theories, as well as potentially generate new ones.

Our work here suggests that semantics is only part of the story. We have focused on evidence suggesting that not all cues *can* be used to symbolize a category. Cues that are too common are useless in signal detection problems, because in virtue of being very common they must necessarily be non-discriminative. We showed that sound symbolism cues are, as this observation predicts, weighted on average in inverse proportion to their logged frequency, i.e., most of the weight is carried by a few rare cues. The requirements of discriminability place constraints on the semantics of sound symbolization in two ways: only a few cues are available to do most of the heavy lifting in sound symbolization, and those can't contradict each other (we can't have the same cue for *small* and *large* things, because then they would be non-discriminative). Even though the constraints made by the need for discriminability are not directly semantic, they delimit the semantics of sound symbolism substantially.

Acknowledgments

This work was made possible by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada to Westbury & Pexman. We thank Ashley Huang for her help on this manuscript, and gratefully acknowledge the helpful advice from our editor and three reviewers on an earlier version. We thank Kristen Deschamps for her help with data collection.

References

Abelin, Å. (2015). Phonaesthemes and sound symbolism in Swedish brand names. *Ampersand, 2*, 19–29.

Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies, 1–4*, 298–348.

Argo, J. J., Popa, M., & Smith, M. C. (2010). The sound of brands. *Journal of Marketing, 74*(4), 97–109.

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex, 63*, 196–205.

Athaide, G. A., & Klink, R. R. (2012). Creating global brand names: The use of sound symbolism. *Journal of Global Marketing, 25*(4), 202–212.

Auracher, J., Albers, S., Zhai, Y., Gareeva, G., & Stavniychuk, T. (2010). P is for happiness, N is for sadness: Universals in sound iconicity to detect emotions in poetry. *Discourse Processes, 48*(1), 1–25.

Aveyard, M. E. (2012). Some consonants sound curvy: Effects of sound symbolism on object recognition. *Memory & Cognition, 40*(1), 83–92.

Baayen, R. H., Milin, P., Filipović-Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438.

Baxter, S. M., Ilicic, J., Kulczynski, A., & Lowrey, T. (2015). Communicating product size using sound and shape symbolism. *Journal of Product & Brand Management, 24*(5), 472–480.

Baxter, S., & Lowrey, T. M. (2011). Phonetic symbolism and children's brand name preferences. *Journal of Consumer Marketing, 28*(7), 516–523.

Bentley, M., & Varon, E. J. (1933). An accessory study of "phonetic symbolism". *The American Journal of Psychology, 45*(1), 76–86.

Berlin, B. (1994). Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nicols, & J. Ohala (Eds.), *Sound symbolism* (pp. 76–93). Cambridge, UK: Cambridge University Press.

Blasi, D. E., Wichmann, S., Hammarstrom, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 10818–10823.

Blumeyer, D. (2012). https://cmloegcmluin.wordpress.com/2012/11/10/relative-frequencies-of-english-phonemes/. Accessed on March 6, 2017.

Bozzi, P., & Flores D'Arcais, G. B. (1967). Experimental research on the intermodal relationships between expressive qualities. *Archivio di Psicologia, Neurologia e Psichiatria, 28*, 377–420.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition, 126*(2), 165–172.

Brown, R., Black, A., & Horowitz, A. (1955). Phonetic symbolism in natural languages. *Journal of Abnormal and Social Psychology, 54*, 312–318.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Cassidy, K. W., Kelly, M. H., & Sharoni, L. A. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology: General, 128*, 362–381.

Coulter, K. S., & Coulter, R. A. (2010). Small sounds, big deals: Phonetic symbolism effects in pricing. *Journal of Consumer Research, 37*(2), 315–328.

Cuskley, C. (2013). Mappings between linguistic sound and motion. *Public Journal of Semiotics, 5*(1), 37–60.

Cuskley, C., Simner, J., & Kirby, S. (2015). Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research Psychologische Forschung*, 1–12.

D'Onofrio, A. (2014). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and Speech, 57*(3), 367–393.

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology, 52*, 259–268.

Derwing, B., Priestly, T., & Westbury, C. (in preparation). *Pronunciation rules for English words: Converting the spelling to sound.*

Derwing, B., & Priestly, T. (1980). *Reading rules for Russian. A systematic approach to Russian spelling and pronunciation*. Columbus, Ohio: Slavica.

Derwing, B. L., Priestly, T. M., & Rochet, B. L. (1987). The description of spelling-to-sound relationships in English, French and Russian: Progress, problems and prospects. *Orthography and Phonology*, 31–52.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass, 6*, 654–672.

Doyle, J. R., & Bottomley, P. A. (2011). Mixed messages in brand names: Separating the impacts of letter shape from sound symbolism. *Psychology & Marketing, 28*(7), 749–762.

Favalli, S., Skov, T., Spence, C., & Byrne, D. V. (2013). Do you say it like you eat it? The sound symbolism of food names and its role in the multisensory product experience. *Food Research International, 54*(1), 760–771.

Fenko, A., Lotterman, H., & Galetzka, M. (2016). What's in a name? The effects of sound symbolism and package shape on consumer responses to food products. *Food Quality and Preference, 51*, 100–108.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America, 102*(2), 1213–1222.

Flumini, A., Ranzini, M., & Borghi, A. M. (2014). Nomina sunt consequentia rerum–sound–shape correspondences with every-day objects figures. *Journal of Memory and Language, 76*, 47–60.

Fort, M., Weiß, A., Martin, A., & Peperkamp, S. (2013). Looking for the bouba-kiki effect in prelexical infants. In: S. Ouni, F. Berthommier & A. Jesse (Eds.) *Proceedings of the 12th international conference on auditory-visual speech processing, August 29–September 1, 2013* (pp. 71–76). Annecy, France.

French, P. L. (1977). Toward an explanation of phonetic symbolism. *Word, 28*, 305–322.

Greenberg, J. H., & Jenkins, J. J. (1966). Studies in the psychological correlates of the sound system of American English. *Word, 22*, 207–242.

Hinton, L., Nichols, J., & Ohala, J. J. (1994). *Sound symbolism*. Cambridge University Press.

Hockett, C. (1963). The problem of universals in language. In J. Greenberg (Ed.), *Universals of language*. Cambridge, MA: MIT Press.

Holland, M. K., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or, *maluma* and *takete* revisited. *Perceptual and Motor Skills, 19*, 111–117.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*(6), 1744–1756.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*(1), 54–65.

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS ONE, 10*(2), e0116494.

International Phonetic Association (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Irwin, F., & Newland, E. (1940). A genetic study of the naming of visual figures. *Journal of Psychology, 9*, 3–16.

Jespersen, O. (1925). *Language: Its nature, development and origin*. New York: Henry Holt & Company.

Johansson, N., & Zlatev, J. (2013). Motivations for sound symbolism in spatial deixis: A typological study of 101 languages. *The Public Journal of Semiotics, 5*, 3–20.

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound symbolism facilitates word learning in English-speaking children. *Cognitive Science, 35*(3), 575–586.

Klink, R. R. (2000). Creating brand names with meaning: The use of sound symbolism. *Marketing Letters, 11*(1), 5–20.

Klink, R. R. (2001). Creating meaningful new brand names: A study of semantics and sound symbolism. *Journal of Marketing Theory and Practice, 9*(2), 27–34.

Klink, R. R. (2003). Creating meaningful brands: The relationship between brand name and brand mark. *Marketing Letters, 14*(3), 143–157.

Klink, R. R., & Athaide, G. A. (2012). Creating brand personality with brand names. *Marketing Letters, 23*(1), 109–117.

Klink, R. R., & Wu, L. (2014). The role of position, type, and combination of sound symbolism imbeds in brand names. *Marketing Letters, 25*(1), 13–24.

Köhler, W. (1929). *Gestalt psychology*. New York, USA: Liveright.

Köhler, W. (1947). *Gestalt psychology* (2nd ed.). New York, USA: Liveright.

Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition, 114*(1), 19–28.

Kuehnl, C., & Mantau, A. (2013). Same sound, same preference? Investigating sound symbolism effects in international brand names. *International Journal of Research in Marketing, 30*(4), 417–420.

LaPolla, R. (1994). An experimental investigation into phonetic symbolism as it relates to Mandarin Chinese. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 130–147). Cambridge, England: Cambridge University Press.

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics, 21*(2), 397–429.

Lockwood, G., & Tuomainen, J. (2015). Ideophones in Japanese modulate the P2 and late positive complex responses. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00933.

Louwerse, M., & Qu, Z. (2017). Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin & Review, 24*(3), 849–855.

Lowrey, T. M., & Shrum, L. J. (2007). Phonetic symbolism and brand name preference. *Journal of Consumer Research, 34*(3), 406–414.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition, 7*(02), 167–193.

Lyons, J. (n.d.) http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/. Accessed on March 6, 2017.

Maglio, S. J., Rabaglia, C. D., Feder, M. A., Krehm, M., & Trope, Y. (2014). Vowel sounds in words affect mental construal and shift preferences for targets. *Journal of Experimental Psychology: General, 143*(3), 1082.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science, 9*(3), 316–322.

Miron, M. S. (1961). A crosslinguistic investigation of phonetic symbolism. *The Journal of Abnormal and Social Psychology, 62*, 623–630.

Miyazaki, M., Hidaka, S., Imai, M., Yeung, H. H., Kantartzis, K., Okada, H., & Kita, S. (2013). The facilitatory role of sound symbolism in infant word learning. *Proceedings of the thirty fifth annual meeting of the cognitive science society*, 3080–3085.

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1152.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369*(1651).

Myers-Schulz, B., Pujara, M., Wolf, R. C., & Koenigs, M. (2013). Inherent emotional quality of human speech sounds. *Cognition & Emotion, 27*(6), 1105–1113.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms <http://www.usf.edu/FreeAssociation/>.

Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology, 45*, 53–75.

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 65*(2), 115.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition, 112*(1), 181–186.

O'Boyle, M. W., Miller, D. A., & Rahmani, F. (1987). Sound-meaning relationships in speakers of Urdu and English: Evidence for a cross-cultural phonetic symbolism. *Journal of Psycholinguistic Research, 16*, 273–288.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica, 40*(1), 1–18.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1–16.

Ohala, J. J., & Eukel, B. W. (1987). Explaining the intrinsic pitch of vowels. In R. Channon & L. Shockey (Eds.), *In honour of ilse lehiste* (pp. 207–215). Dordrecht: Foris.

Ohtake, Y., & Haryu, E. (2013). Investigation of the process underpinning vowel-size correspondence. *Japanese Psychological Research, 55*(4), 390–399.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology, 114*(2), 173–186.

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

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research, 214*(3), 373–380.

Parise, C., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research, 220*, 319–333.

Park, J. W., & Osera, S. (2008). The effect of brand sound on consumers' brand evaluation in Japan. *Japan Association for Consumer Studies, 16*(2), 23–36.

Paulesu, E., Harrison, J., Baron-Cohen, S., Watson, J. D., Goldstein, L., Heather, J., ... Frith, C. D. (1995). The physiology of coloured hearing. A PET activation study of colour-word synaesthesia. *Brain, 1118*(3), 661–676.

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science, 22*(11), 1419–1421.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE, 10*(9), e0137147.

Plato (360BCE/1892). *Cratylus*. Translated by Benjamin Jowett <http://www.gutenberg.org/cache/epub/1616/pg1616.txt>. June 27, 2016.

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

Reilly, J., Hung, J., & Westbury, C. (2017). Non-Arbitrariness in mapping word form to meaning: Cross-linguistic formal markers of word concreteness. *Cognitive science, 41*(4), 1071–1089.

Reilly, J., Westbury, C., Kean, J., & Peele, J. (2012). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PLoS ONE, 7*(8), e42286. https://doi.org/10.1371/journal.pone.0042286.

Rhodes, R. (1994). Aural images. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 276–292). Cambridge University Press.

Roblee, L., & Washburn, M. F. (1912). The affective values of articulate sounds. *The American Journal of Psychology, 23*(4), 579–583.

Rogers, S., & Ross, A. (1975). A cross-cultural test of the maluma–takete phenomenon. *Perception, 5*(2), 105–106.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*(3), 225–239.

Saussure, F. (1916/1983) Course in general linguistics. In Charles Bally, Albert Sechehaye (Eds,). *Trans. Roy Harris*. La Salle, Illinois: Open Court.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423.

Shaoul, C. & Westbury, C. (2006). *USENET orthographic frequencies for 111,627 English words (2005–2006)*. Edmonton, AB: University of Alberta <http://www.psych.ualberta.ca/~westburylab/downloads/wlfreq.download.html>.

Shillcock, R., Kirby, S., McDonald, S., & Brew, C. (2001). Filled pauses and their status in the mental lexicon. In *Proceedings of the 2001 conference of disfluency in spontaneous speech* (pp. 53–56). Edinburgh: International Speech Communication Association.

Shinohara, K., & Kawahara, S. (2010). A cross-linguistic study of sound symbolism: The images of size. In *Proceedings of the 36th annual meeting of the Berkeley Linguistics Society* (36, pp. 396–410). Berkeley Linguistics Society.

Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? Sound symbolism and gender in first names. *PLoS ONE, 10*(5), e0126809.

Sidhu, D. M., & Pexman, P. M. (2016). A prime example of the maluma/takete effect? Testing for sound symbolic priming. *Cognitive Science*. https://doi.org/10.1111/cogs.12438.

Sidhu, D. M., & Pexman, P. M. (2017). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1361-1.

Sidhu, D. M., Pexman, P. M., & Saint-Aubin, J. (2016). From the Bob/Kirk effect to the Benoit/Éric effect: Testing the mechanism of name sound symbolism in two languages. *Acta Psychologica, 169*, 88–99.

Simner, J., Cuskley, C., & Kirby, S. (2010). What sound does that taste? *Perception, 39*, 553–569.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995. https://doi.org/10.3758/s13414-010-0073-7.

Tanz, C. (1971). Sound symbolism in words relating to proximity and distance. *Language and Speech, 14*, 266–276.

Tarte, R. D. (1982). The relationship between monosyllables and pure tones: An investigation of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior, 21*, 352–360. https://doi.org/10.1016/S0022-5371(82)90670-3.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83.

Ultan, R. (1978). Size-sound symbolism. In Joseph H. Greenberg (Ed.), *Universals of human language. Phonology* (Vol. 2). Stanford: Stanford University Press.

Usnadze, D. (1924). Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung. *Psychological Research Psychologische Forschung, 5*(1), 24–43.

Vainio, L., Schulman, M., Tiippana, K., & Vainio, M. (2013). Effect of syllable articulation on precision and power grip performance. *PLoS ONE, 8*. https://doi.org/10.1371/journal.pone.0053061.

Von Humboldt, W. (1836). *On language: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge, UK: Cambridge University Press.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnsson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*(1), 21–25.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain & Language, 93*(1), 10–19.

Westbury, C., Hollis, G., & Shaoul, C. (2007). LINGUA: The language-independent neighbourhood generator of the University of Alberta. *The Mental Lexicon, 2*, 271–284.

Westbury, C., Shaoul, C., Moroschan, G., & Ramscar, M. (2016). Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language, 86*, 141–156.

Westbury, C. (In press). Implicit sound symbolism effects in lexical access, revisited: A requiem for the interference task paradigm. Accepted for publication. In: Journal of Articles in Support of the Null Hypothesis.

Yorkston, E., & Menon, G. (2004). A sound idea: Phonetic effects of brand names on consumer judgments. *Journal of Consumer Research, 31*(1).