Journal of Memory and Language 125 (2022) 104323

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Higher order factors of sound symbolism

David M. Sidhu [a,*], Gabriella Vigliocco [b], Penny M. Pexman [a]

[a] University of Calgary, Canada
[b] University College London, UK

ABSTRACT

Sound symbolism refers to associations between certain language sounds (i.e., phonemes) and perceptual and/or semantic properties. Crucially, the different associations of a phoneme do not appear to be wholly independent. For instance, the phoneme /i/ is associated with sharpness, smallness and brightness. Previous work has shown that these properties are all related to one another (Walker et al., 2012). This suggests that higher order factors may underlie sound symbolic associations. In Experiment 1 we measured 25 different associations of phonemes and found that these associations clustered according to the higher order factors of: activity, valence, potency and novelty. In addition, certain phonemes were found to go along with different higher order factors. Then, in Experiments 2a and 2b, we demonstrated that higher order factors can play a role in associations between phonemes and abstract shape stimuli. Together these results characterize the role of higher order semantic properties in sound symbolism and contribute to our understanding of the mechanisms underlying sound symbolism.

Introduction

Sound symbolism refers to associations between phonemes and particular perceptual and/or semantic properties (see Sidhu & Pexman, 2018). That is, there is evidence that phonemes are inherently associated with certain kinds of things. The most well-known example of this is the maluma/takete effect (Köhler, 1929) in which certain phonemes (e.g., those in *maluma*) seem associated with round and smooth shapes, and others (e.g., those in *takete*) seem associated with sharp and spiky shapes. In general, it seems that sonorants (e.g., /l/, /m/, /n/), voiced stops (e.g., /b/, /d/, /g/) and back/rounded vowels (e.g., /oʊ/ as in *boat*) are associated with roundness; while voiceless stops (e.g., /p/, /t/, /k/) and front/unrounded vowels (e.g., /i/ as in *beet*) are associated with sharpness (McCormick et al., 2015; see also Knoeferle et al., 2017). This has typically been demonstrated by asking participants to pair nonwords like *maluma* and *takete* with a round and a sharp shape (see Fig. 1), in the way that seems most natural. Roughly 90% of participants (Styles & Gawne, 2017) tend to pair nonwords and shapes in a way that is congruent with the maluma/takete effect (for variation based on language see Styles & Gawne, 2017; Cwiek et al., 2022; based on age see Fort et al., 2018; Pejovic & Molnar, 2016). Thus, something in the sound, articulation and/or visual properties (including orthography and/or mouth movements) of these phonemes leads to an association with roundness or sharpness (see Sidhu & Pexman, 2018).

While shape sound symbolism is the most prominent example of sound symbolism, it is by no means the only one. Another example is size sound symbolism (i.e., the mil/mal effect; Sapir, 1929) in which high-front vowels (e.g., /i/) show an association with small shapes, and low-back vowels (e.g., /ɑ/ as in *bought*) show an association with large shapes. Beyond shape and size, sound symbolic associations have been demonstrated for the dimensions of speed (Cuskley, 2013), personality (Sidhu et al., 2019), brightness (Newman, 1933), arousal (Aryani et al., 2018), taste (Gallace et al., 2011), social dominance (Auracher, 2017), and colour (Kim et al., 2018), to name a few.

Various mechanisms have been proposed for sound symbolic associations (reviewed in Sidhu & Pexman, 2018). One is that they arise from a co-occurrence among sounds and features in the world. For example, the reason that high-front vowels are associated with smallness may be that these vowels tend to have a higher pitch, and that smaller things in the real world tend to resonate at a higher pitch (see Spence, 2011). Another proposed mechanism is that phonemes and associated features might share some property in common; in particular, a property that can exist across modalities. For instance, Aryani et al. (2020) recently demonstrated that nonwords like *takete* elicit high levels of affective arousal, as do sharp shapes. They suggest that this shared property might contribute to the maluma/takete effect.

One way of identifying such shared properties is to examine commonalities among the various associations of specific phonemes. Indeed, surveying the existing work on sound symbolism suggests that the various associations of a phoneme are not entirely distinct phenomena. Rather, there seem to be patterns in the associations of a given phoneme. French (1977) noted that the various associations of the phoneme /i/ are

* Corresponding author at: Department of Experimental Psychology, University College London. Experimental Psychology, University College London, Gower Street WC16BT, London, UK.

E-mail address: dmsidhu@gmail.com (D.M. Sidhu).

https://doi.org/10.1016/j.jml.2022.104323
Received 31 March 2021; Received in revised form 9 February 2022; Accepted 25 February 2022
Available online 12 March 2022

related to one another. For example, /i/ is associated with: smallness, brightness, and sharpness. These dimensions are connected, for instance, in the fact that small objects are rated as having associations with brightness and sharpness (Walker et al., 2012). Similar patterns can be observed for consonants. Compared to voiced stops, voiceless stops are more associated with smallness, quickness and sharpness (Klink, 2000); and these dimensions are also related to one another (Walker et al., 2012).

The main goal of the present work was to examine the interrelationships among a large number of sound symbolic associations, in order to identify the higher order properties that they have in common. We then measured phonemes' associations with each of these higher order properties in order to explore the role of higher order factors in sound symbolism. In this way, we moved beyond examining local associations between phonemes and specific dimensions, to a more global approach in which we examine associations between phonemes and groups of dimensions, and the higher order factors that unite them. In the remaining sections of the Introduction we review previous work exploring higher order properties of semantics in general, and then in sound symbolism specifically.

Higher order semantic properties

Given our goal, we decided to use the semantic differential approach (i.e., extracting latent dimensions from ratings on bipolar dimensions, Osgood, et al., 1957). Osgood et al had participants rate various words on semantic differential scales: scales anchored by antonyms (e.g., a seven-point scale anchored by "pleasant" on one end, and "unpleasant" on the other). Their factor analysis of these ratings suggested three underlying factors. The first was related to the overall pleasantness of a concept, defined by scales such as: *good-bad* or *pleasant-unpleasant.* Osgood et al. originally termed this an *evaluative* factor, but it has since come to be referred to as *valence* (e.g., Warriner et al., 2013). The second factor was termed *potency* and had to do with the overall "toughness" (p. 63) of the concept. Hollis and Westbury (2016) characterized it as "the degree to which [the concept] could affect change" (p. 1744). It was defined by scales such as: *strong–weak* and *rugged-delicate*. The final factor was termed *activity*, characterized by Hollis and Westbury (2016) as the concept's "energetic potential" (p. 1744). It was defined by scales such as: *fast-slow* and *sharp-dull*. Note that these factors were not entirely orthogonal. For instance, high ends of the potency and activity factors tend to have a positive valence. It is important to mention that other factors emerged in the studies conducted by Osgood et al. (1957). For instance, a *stability* factor, defined by dimensions such as *stable-changeable* and *orthodox-heretical*, also emerged in several analyses. In the years since Osgood et al. (1957) a number of other factors have been discovered. These include factors defined as *orderliness*, *reality*, *familiarity*, *and complexity* (Bentler & LaVoie, 1972; Malhotra, 1981; Trofimova, 2014; Wickens & Lindberg, 1975).

We decided to use this approach for the following reasons. First, it has been shown that higher order dimensions extracted by this method (i.e., valence, activity and potency) generalize to various stimulus types, including paintings, sculptures, sonar signals (Osgood et al., 1957) and colours (Fang et al., 2015). In addition, dimensions such as valence and activity (more recently termed arousal) have been found to affect language processing (e.g., Estes & Adelman, 2008; Kuperman et al., 2014; Kousta et al., 2009; Vinson et al., 2014). This indicates that the dimensions identified with such an approach have validity. Second, collecting ratings on bipolar dimensions leads to higher order dimensions that are easily interpretable. While others have extracted latent dimensions from, for example, word co-occurrence vectors (e.g., Hollis & Westbury, 2016) the resulting factors are more often opaque, which does not serve the present purpose. Finally, as described previously, the use of contrasting bipolar dimensions is consistent with how research on sound symbolism has typically been done. The vast majority of work on sound symbolism has involved nonword decisions that are anchored by a pair of contrasting stimuli (e.g., contrasting shapes, contrasting sensations), either as a binary choice or a rating scale. Indeed, it is difficult to imagine another way to go about collecting the explicit sound symbolic associations of a nonword. Having participants generate features or associates themselves would likely result in very heterogenous data. We should also note that we are not arguing in favour of a particular theory of semantic representation—there are certainly other ways of conceptualizing semantics (e.g., a featural approach to meaning; McRae et al., 1997; Vigliocco et al., 2004). However, the semantic differential approach was the best way to characterize the phenomenon that we sought to explain.

Several studies have applied the semantic differential technique to nonwords, allowing an examination of the factors underlying sound symbolic associations. Miron (1961), a student of Osgood's, reported the general finding that more anterior consonants (i.e., those articulated at the front of the mouth; e.g., /p/) and front vowels were judged as more positive in valence, and lower in potency. However, this was based on observations of trends rather than formal analyses. Another example is the largescale study conducted by Greenberg and Jenkins (1966). In several experiments, they had participants rate individual consonants and vowels on 26 different semantic differential scales. Their general result for consonants suggested a three-factor structure. The first was characterized as a distinction between concentration and dispersion, including scales such as: *abrupt-continuous*, and *liquid–solid*. The second factor was characterized as a distinction between harshness and mellowness, the third was identified as a potency factor. The authors did not statistically test the alignment of different consonants with each factor. However, they observed a trend in which stops tended to fall at

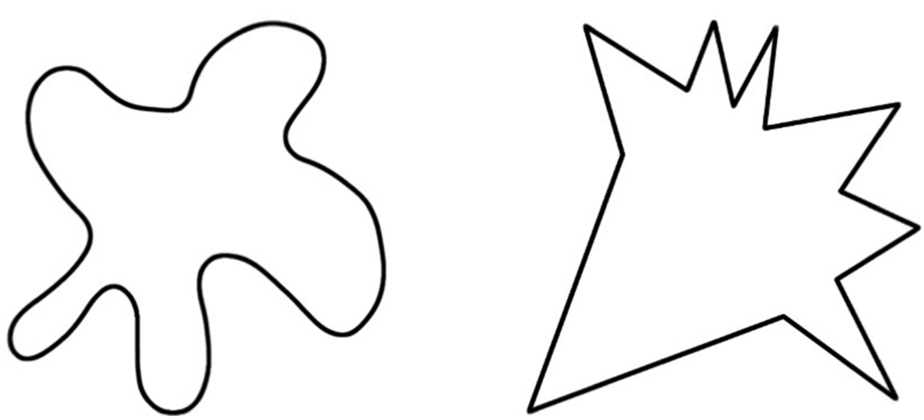

Fig. 1. *Prototypical shapes used in maluma/takete matching tasks. Note.* Participants typically pair nonwords like *maluma* with the round shape on the left, and nonwords like *takete* with the sharp shape on the right.

one end of the factors (the concentrated, harsh, and strong ends); and nasals, sibilants and /l/ tended to fall at the other. Their analyses of vowels generally revealed three factors as well. They defined the first as a distinction between acuteness and graveness, including scales such as: *high-low*, *sharp-dull*, *thick-thin.* The second factor corresponded to valence, and the third was a combination of potency and expansiveness. See Table 1 for a summary of the higher order semantic properties mentioned. They also observed a general tendency for front and back vowels to appear at opposite ends of these factors (with front vowels at the acute, unpleasant and weak ends). This suggests that there are different higher order semantic dimensions associated with different phoneme types. However, there is a need to examine this pattern statistically.

A more recent effort to uncover higher order properties in sound symbolism was conducted by Tzeng, Nygaard, and Namy (2017). Instead of using a semantic differential technique, they had participants guess the meanings of foreign words. Meanings consisted of four antonyms describing the dimensions of size, shape, speed and movement. These meanings were then rated on a variety of scales representing potential higher order dimensions (including Osgood et al.'s factors of intensity, concreteness and magnitude). The authors then examined whether ratings on these scales explained participant responses. They found some evidence that the higher order dimension of *intensity* could contribute to ratings but concluded that phoneme-dimension pairings tended to be separate, at least for the four dimensions studied.

Finally, a study by Westbury et al. (2018) took a big data approach and examined the fit between nearly 8000 nonwords and six dimensions: size, shape, gender, valence and concreteness. Interestingly, the fits between nonwords and several of the semantic categories were correlated. For example, nonwords that were judged as a good fit for something *large* were also judged as good fits for something *round* and *feminine*. This again suggests relationships among sound symbolic associations.

Present study

Previous work suggests relationships among various sound symbolic associations of certain phonemes. However, while studies using the semantic differential technique sampled a broad range of dimensions, they tended not to test the relationships between phonemes and the higher order factors that they extracted. For instance, Greenberg and Jenkins (1966) did not test for an association between specific phoneme categories and the higher order factors they extracted. More recent studies have used sophisticated analysis techniques but have only examined a small number of dimensions. In the present study, our main goal was to statistically explore the interrelationships among many sound symbolic associations. In particular, here we examined a large number of sound symbolic associations, identified their interrelationships, and characterized the higher order factors underlying those interrelationships. We then explored the associations between phoneme categories (e.g., voiceless stops) and these higher order factors.

Table 1
List of higher order semantic properties mentioned in Introduction.

Higher Order Property	Example Dimensions
Valence	*good-bad; pleasant-unpleasant*
Arousal	*active–passive; sharp-dull*
Potency	*strong–weak; rugged-delicate*
Stability	*stable-changeable; orthodox-heretical*
Orderliness	*structured-disorganized; orderly-disarrayed*
Reality	*authentic-fake; concrete-abstract*
Familiarity	*commonplace-exceptional; regular-rare*
Complexity	*complex-simple; mysterious-usual*
Concentration	*abrupt-continuous; solid–liquid*
Harshness	*harsh-mellow; rough-smooth*
Acuteness	*acute-grave; narrow-wide*

In Experiment 1 we collected nonword ratings on 25 semantic differential scales (described below). Based on these ratings we addressed several questions about the involvement of higher order semantic properties in sound symbolism:

1) What are the associations of different phoneme categories?
2) Are there higher order factors that can be observed in phonemes' associations?
3) What are the associations between different phoneme categories and the higher order factors extracted?

Then, as a secondary investigation, we explored the extent to which these higher order factors play a role in the associations between nonwords and perceptual stimuli. Thus, in Experiment 2a, we collected ratings of abstract images on the same 25 semantic dimensions and measured the associations between nonwords and these abstract images using a rating scale, followed by a forced judgment task (Experiment 2b). This allowed us to address a final question:

4) Do these higher order factors explain the fit between nonwords and visual stimuli?

Experiment 1

Method

Participants

Participants were 104 undergraduate students at the University of Calgary who participated in exchange for course credit. All participants were fluent in English, reported normal or corrected to normal vision and provided informed consent.

Materials

Stimuli consisted of 40 CVCV nonwords. This syllable structure is consistent with McCormick et al. (2015) and was also chosen so that there would be an equal number of consonants and vowels in each nonword. Each nonword contained two different consonants with the same manner of articulation, either sonorants (/l/, /m/, /n/), voiceless stops (/p/, /t/, /k), voiced stops (/b/, /d/, /g/), voiceless fricatives (/f/, /s/, /ʃ/) or voiced fricatives (/v/, /ð/, /z/). There were eight nonwords containing each kind of consonant. Within each group of eight, half of the nonwords contained two front vowels (one each of /i/ and /eɪ/) and half contained two different back vowels (one each of /ɑ/ and /oʊ/). With these rules in mind, there were twelve possible nonwords for each consonant–vowel combination. From these twelve possibilities we chose four somewhat arbitrarily, while attempting to balance inward vs outward patterns of articulation (see Topolinski et al., 2014) and which phonemes appeared first in the nonwords. We also ensured that nonwords were phonotactically legal in English and were not homophones of existing words. This was confirmed by a trained linguist.

The 40 nonwords consisted of two lists (henceforth *List A* and *List B*), containing an equal number of each nonword type.[1] A female psycholinguist blind to the purpose of the study recorded each of the nonwords in List A with a flat intonation. A female voice actress also blind to the purpose of the study recorded the nonwords in List B in a similar manner. We ensured that the average pitch of nonword recordings was as similar as possible ($Range_{List\ A}$ = 181.79–197.19 Hz; $SD_{List\ A}$ = 3.98 Hz; $Range_{List\ B}$ = 168.35–185.26 Hz; $SD_{List\ B}$ = 4.83 Hz).

Nonword stimuli were rated on 25 semantic dimensions. Our goal was to sample as broad a range of dimensions as could be motivated from the previous literature on semantics and/or sound symbolism. Thus, we included three dimensions for each of the three factors

[1] Ratings for each list were collected as separate studies, explaining the different speakers and settings (i.e., in person vs. online).

discovered by Osgood et al. (1963): *good-bad*, *beautiful-ugly*, *pleasant-unpleasant* (representing valence); *strong–weak*, *big-small*, *rugged-delicate* (representing potency); and *active–passive*, *fast-slow*, *sharp-round* (representing activity). We also included dimensions to represent other factors that have been found since (Bentler & LaVoie, 1972; Malhorta, 1981; Trofimova, 2014; Wickens & Lindberg, 1975): *realistic-fantastical*, *structured-disorganized*, *ordinary-unique*, *interesting-uninteresting*, and *simple-complex*. Next we included dimensions that previous studies of sound symbolism and/or crossmodal correspondences (general associations between stimulus dimensions; e.g., size and pitch) have found to be relevant (Greenberg & Jenkins, 1966; Miron, 1961; Sidhu & Pexman, 2015; Tarte, 1982; Walker et al., 2012): *abrupt-continuous*, *exciting-calming*, *hard-soft*, *happy-sad*, *harsh-mellow*, *heavy-light*, *inhibited-free*, *masculine-feminine*, *solid-nonsolid*, and *tense-relaxed*. Finally, we included *dangerous-safe* as a dimension that has been shown to be important to word meaning (Wurm, 2007). Note that many of these dimensions were relevant for multiple reasons (e.g., *sharp-round* represents the activity factor, and is also a key dimension for sound symbolism).

Procedure

Of the 104 participants, 58 took part in person and made their ratings in our laboratory (those rating List A) while 46 took part online (those rating List B). Both versions of the task used surveys hosted by the survey platform Qualtrics. Participants rated a random 15 nonwords from their list, one at a time. We elected to present each participant with 15 nonwords because feedback from pilot participants suggested that this number of nonwords could be rated without participants becoming fatigued. A sound file for each nonword was presented at the top of each page. Participants could play this as many times as they wished. They then rated that nonword on the 25 dimensions. The instructions emphasized that participants were to rate the *impression* of the nonword:

> You will hear fifteen nonwords (made up words that don't mean anything), one at a time, and be asked to rate each of them on a variety of different scales. We want you to rate these nonwords based on the impression that you get from them. So, even though they don't mean anything, rate them based on the general *impression* you get from them.
> Importantly, some of these ratings will not be very literal. For instance, imagine that you were asked to rate a nonword on a scale from warm to cold. This would be difficult to do literally. However, you would be able to rate the nonword based on whether its sound gives off a warm or a cold *impression*.
> There are no right or wrong answers; we are interested in what - you feel to be the best answer.
> Don't spend too long on any particular rating; try to go with your first instinct.

Rating dimensions were presented as seven-point scales anchored by each adjective. Nonwords and dimensions were presented in a random order. The online version of this task also included an attention check item (i.e., a sound file asking the participant to select seven for each scale). Participants then completed a debriefing questionnaire which asked them if they had any problem focusing, if any of the nonwords were real words in a language they spoke, and whether their data should be used (online study only).

Results

Data analysis

We used different approaches to address each of the first three questions outlined in the introduction. To answer **"1) What are the associations of different phoneme categories?"**, we used regression models predicting the ratings of different phoneme categories on each of the 25 semantic differential scales. Next, to address **"2) Are there higher order factors that can be observed in phonemes' associations?"**, we ran exploratory factor analyses to quantify the structure of associations, and to identify higher order latent variables (i.e., factors) among the associations. Finally, in order to address **"3) What are the associations between different phoneme categories and the higher order factors observed in Question 2?"**, we used regression models predicting the scores of different phoneme categories on the factors extracted by the factor analysis. Data and code for all analyses can be found at https://osf.io/gruqs/.

Data cleaning

We excluded participants who gave the same response to each scale for more than two nonwords (three participants), said that they could not focus (two participants), research assistants reported issues with (in person only; 12 participants)[2], failed the attention check (online only; four participants), or told us not to use their data (online only; six participants). These were not mutually exclusive, and in total the data for 19 participants were removed. We also removed trials for nonwords that participants reported were real words in a language they spoke (removed on a participant-by-participant basis; 13 trials). Due to a programming error, 22 participants received the wrong audio file for the nonword *neelay*. These trials were also removed.

We examined the reliability of these ratings by calculating the Intraclass Correlation Coefficient (ICC2k) for each dimension using the "psych" package in R (Revelle, 2021). All but two dimensions showed good reliability (>.75). The interesting-uninteresting dimension showed only moderate reliability (.69) while the structured-disorganized dimension showed low reliability (.39). Because of this the structured-disorganized dimension was removed from all further analyses with nonwords. See Table S1 for each dimension's ICC2k value.

What are the associations of different phoneme categories?

We used linear mixed effects models to examine whether categories of phonemes differed in their association with any of the 24 dimensions. For each rating dimension, we computed a model using rating as the dependent variable. Models included vowel type (effects coded; front vowels [-.5] and back vowels [.5]) and consonant type (dummy coded; sonorants as the reference category). In addition, models included random subject slopes for vowel and consonant type, as well as random subject and item intercepts. Due to convergence issues, we did not include correlations between slopes and intercepts. In cases where convergence was still not achieved, or a singular fit was returned, random slopes were removed beginning with the one with the least amount of variance, until a good fit was found. See Table 2 for the marginal R^2 of each model, calculated using the "MuMIn" package (Bartón, 2020).

After computing a model for each rating scale, we used the "emmeans" package in R (Lenth, 2018) to compare the estimated marginal mean of each consonant category to the overall mean of that rating scale. We corrected for multiple comparisons using false discovery rate correction. Significant effects ($p < .05$) are shown in Fig. 2. See Fig. 3 for significant effects of vowel category. In Fig. 4 we present nonword ratings on several dimensions of interest. See Figures S1 and S2 in Supplementary Material for comparisons to each participant's midpoint, rather than the overall mean.

In order to test the robustness of these effects to different lists, voices, and testing contexts (i.e., in person vs. online), we ran versions of these models including an interaction between consonant category and list. We followed up the significant interactions and found that six of the sixty consonant effects only emerged when examining List A. These were sonorants' association with ordinary-unique, voiced fricatives'

[2] In order to obtain high quality ratings, we erred on the side of caution and removed participants for reasons such as participants not reading instructions, excess noise in the lab at the time, and technical issues. These exclusions were made prior to beginning the analyses.

Table 2
The marginal R^2 of fixed effects (i.e., consonant and vowel type) in the prediction of each rating scale.

Scale	Marginal R^2
Abrupt-Continuous	0.078
Beautiful-Ugly	0.042
Big-Small	0.097
Dangerous-Safe	0.051
Delicate-Rugged	0.094
Exciting-Calming	0.110
Fast-Slow	0.084
Good-Bad	0.013
Happy-Sad	0.023
Hard-Soft	0.094
Harsh-Mellow	0.085
Heavy-Light	0.069
Inhibited-Free	0.036
Interesting-Uninteresting	0.015
Masculine-Feminine	0.124
Ordinary-Unique	0.030
Passive-Active	0.058
Pleasant-Unpleasant	0.024
Realistic-Fantastical	0.031
Sharp-Round	0.141
Simple-Complex	0.079
Solid-NonSolid	0.054
Strong-Weak	0.053
Tense-Relaxed	0.059

Note. Marginal R2 was calculated using the approach described in Nakagawa and Schielzeth (2013), and Nakagawa et al. (2017). R^2 values for mixed models are an approximation, and not exactly equal to proportion of variance explained. They are presented here to allow comparison across different scales.

associations with abrupt-continuous and hard-soft, voiceless fricatives' association with hard-soft, and voiceless stops' associations with abrupt-continuous and hard-soft. This may have been due to better audio quality for the study run in the lab as opposed to online.

Are there higher order factors that can be observed in phonemes' associations?

The main goal of this paper was to examine the higher order factors that emerge in sound symbolism. Before proceeding to a factor analysis, we conducted a network analysis to visualize the relationships between different sound symbolic associations. This was done using the "qgraph" (Epskamp & Fried, 2018), "bootnet" (Epskamp et al., 2015), and "BGGM" (Williams & Mulder, 2019) packages in R. In general, network analysis involves computing associations among pairs of variables, while accounting for all other variables in the network. We refer the reader to Epskamp and Fried (2018) for a fuller description of the process. Here we computed networks using two different approaches that have been shown to have either high sensitivity (low Type 2 error) or specificity (low Type 1 error; based on simulations by Isvoranu & Epskamp, 2021). Dimensions were mean-centered within participants to account for non-independence (see Costantini et al., 2019), and non-paranormal transformations were used to account for skewed data (see Isvoranu & Epskamp, 2021). The resulting networks can be seen in Fig. 5.

We then computed a centrality index for each dimension, which is a measure of how interconnected a given dimension is to all others. Dimensions scoring higher on this measure are connected to, and affect ratings on, a greater number of dimensions. This is defined as the sum of absolute partial correlation coefficients for that dimension, see Fig. 6. Note that simulations suggested our high sensitivity approach was the most accurate at identifying centrality (Isvoranu & Epskamp, 2021), and so we calculated centrality based on this network. Delicate-rugged was the most central dimension (1.27), and significantly more central than 17 of the 24 dimensions ($p < .05$). Harsh-mellow (1.24) and hard-soft (1.23) were the next most central dimensions, and significantly more central than 14 of the 24 dimensions (p's $< .05$).

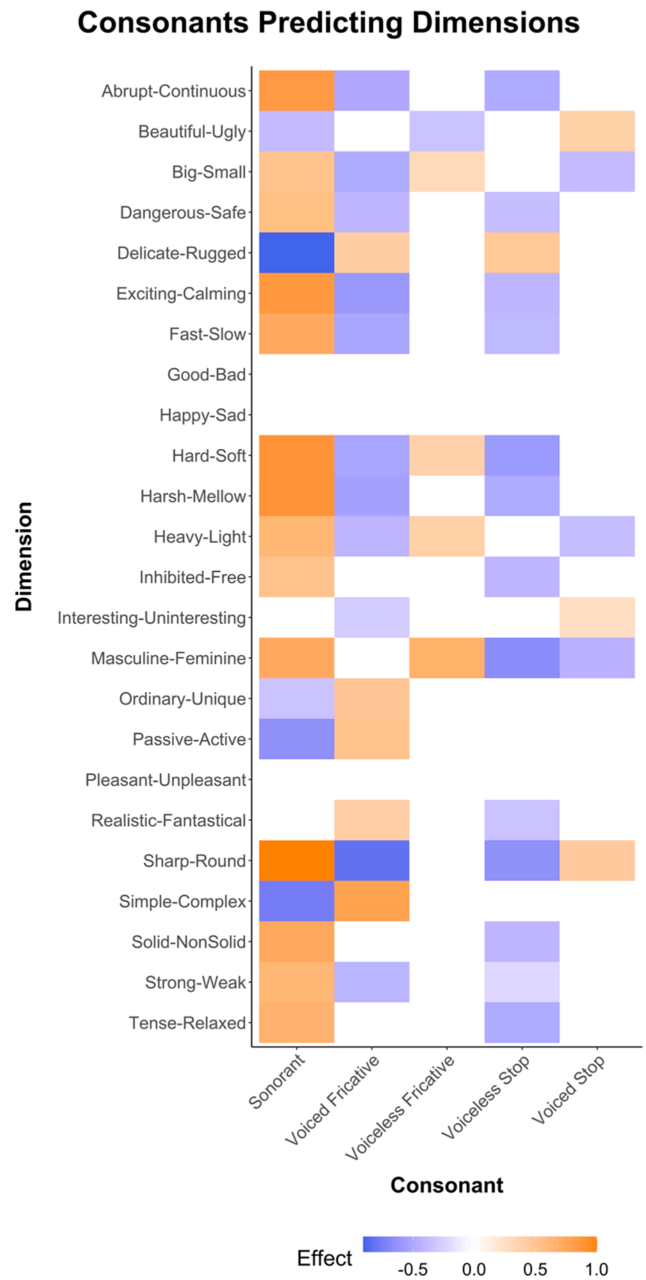

Fig. 2. Results of consonant categories predicting dimension scores. In particular, the differences between the estimated marginal means for each consonant category and overall mean are shown. *Note.* Warm (cool) colours indicate that a consonant category was associated with the end of the dimension denoted by the second (first) term, compared to the mean across all nonwords. Only significant effects are shown ($p < .05$, FDR correction applied for five tests).

Nonword ratings were then submitted to an exploratory factor analysis. We used the "psych" package in R (Revelle, 2021) to conduct a parallel analysis determining the optimal number of factors to extract. This process involves bootstrapping datasets from the observed data, and then computing eigenvalues for both the observed data and bootstrapped datasets. The number of factors to extract is equal to the number of factors for which eigenvalues in the observed data are greater than those in the bootstrapped samples. This approach suggested a four-factor structure. We used principal axis factoring because our data did not display multivariate normality (see Osborne & Costello, 2005). We began with an oblimin rotation. We did not conduct an orthogonal rotation because several correlations among factors were greater than r

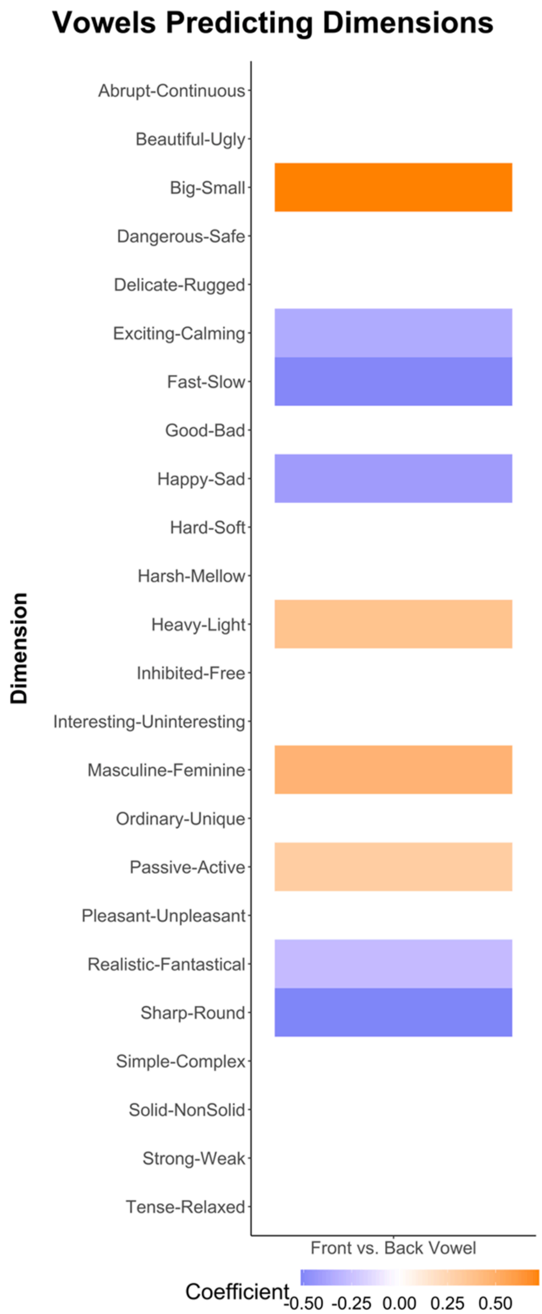

Fig. 3. Results of vowel type predicting dimension scores. In particular, coefficients for the vowel type predictor are shown. Note. Warm (cool) colours indicate that front (back) vowels were associated with the end of the dimension denoted by the second term. Only significant effects are shown ($p < .05$).

= .30 (see Myers et al., 2012). The fit statistics for our factor analysis suggested a good fit (RMSEA = 0.03, TIL = 0.97).

See Table 3 for factor loadings. The first three factors seem to correspond to the three factors from Osgood et al. (1957), namely activity, valence and potency. We have named the final factor novelty based on the two dimensions with the highest loadings: *ordinary-unique* and *realistic-fantastical.* To examine the robustness of this analysis to different lists, we performed this factor analysis separately on Lists A and B. Both analyses resulted in a four-factor solution consisting of the same four factors (see Table S1 in Supplementary Material for each factor loadings). See Table 4 for the correlations among factors.

What are the associations between different phoneme categories and the extracted higher order factors?

We next examined the relationships between different phoneme types and the factors extracted by the exploratory factor analysis. To that end, we conducted linear mixed effects regressions predicting extracted factor scores, in the same manner as reported previously for individual dimensions. As before, we compared the estimated marginal mean of each consonant category to the overall mean of a factor. The results are shown in Fig. 7. Sonorants were associated with low activity (Difference Between Estimated Marginal Mean and Overall Mean [*EMMD*] = -0.60, $p < .001$), while voiced fricatives (*EMMD* = 0.39, $p < .001$) and voiceless stops were associated with high activity (*EMMD* = 0.36, $p < .001$). Sonorants (*EMMD* = -0.49, $p < .001$) and voiceless fricatives (*EMMD* = -0.26, $p = .004$) were associated with low potency, while voiced fricatives (*EMMD* = 0.28, $p = .003$), voiceless stops (*EMMD* = 0.23, $p = .007$) and voiced stops (*EMMD* = 0.24, $p = .005$) were associated with high potency. Sonorants (*EMMD* = -0.26, $p < .001$) and voiced stops (*EMMD* = -0.17, $p = .01$) were associated with low novelty, while voiced fricatives (*EMMD* = 0.40, $p < .001$) were associated with high novelty. See Online Supplementary Material for further analyses involving phoneme sonority, voicing and manner of articulation. Front (back) vowels were associated with high (low) activity ($b = 0.24$, $p = .01$) and low (high) potency ($b = -0.25$, $p = .003$).

Discussion

By examining a large number of sound symbolic associations, we were able to characterize their interrelationships. In particular, we found that associations grouped according to the higher order factors of activity, valence, potency and novelty. Importantly, we found associations between phoneme categories and these latent factors: high (low) activity was associated with voiced fricatives, voiceless stops and front vowels (sonorants and back vowels); high (low) potency was associated with voiced fricatives, stops and back vowels (sonorants, voiceless fricatives and front vowels); high (low) novelty was associated with voiced fricatives (sonorants and voiced stops).

Experiment 2a

Our next goal was to address the question **"4) Do higher order factors explain the fit between nonwords and visual stimuli?"**. Various studies have shown that phonemes have associations with perceptual stimuli, such as shapes (e.g., Nielsen & Dingemanse, 2011) or tastes (Gallace et al., 2011). Our goal was to examine whether shared higher order factors between nonwords and non-auditory perceptual stimuli play a role in these associations. To that end, we collected ratings for abstract images on the same dimensions on which the nonwords were rated and subjected these to factor analysis. To presage the use of these ratings, in the current experiment we examined whether similarities in higher order factors associated with nonwords and images contributed to their rated fit with one another. In Experiment 2b we tested whether participants will choose an image that is highly similar to a given nonword on higher order factors (vs. highly dissimilar) as the better fit for that nonword.

Image ratings

Method

Participants. Participants were 53 undergraduate students at the University of Calgary who participated in exchange for course credit. All participants were fluent in English, reported normal or corrected to normal vision and provided informed consent.

Materials. Materials consisted of 20 abstract shapes. These were freely available from the Noun Project (https://thenounproject.com/) and Flat

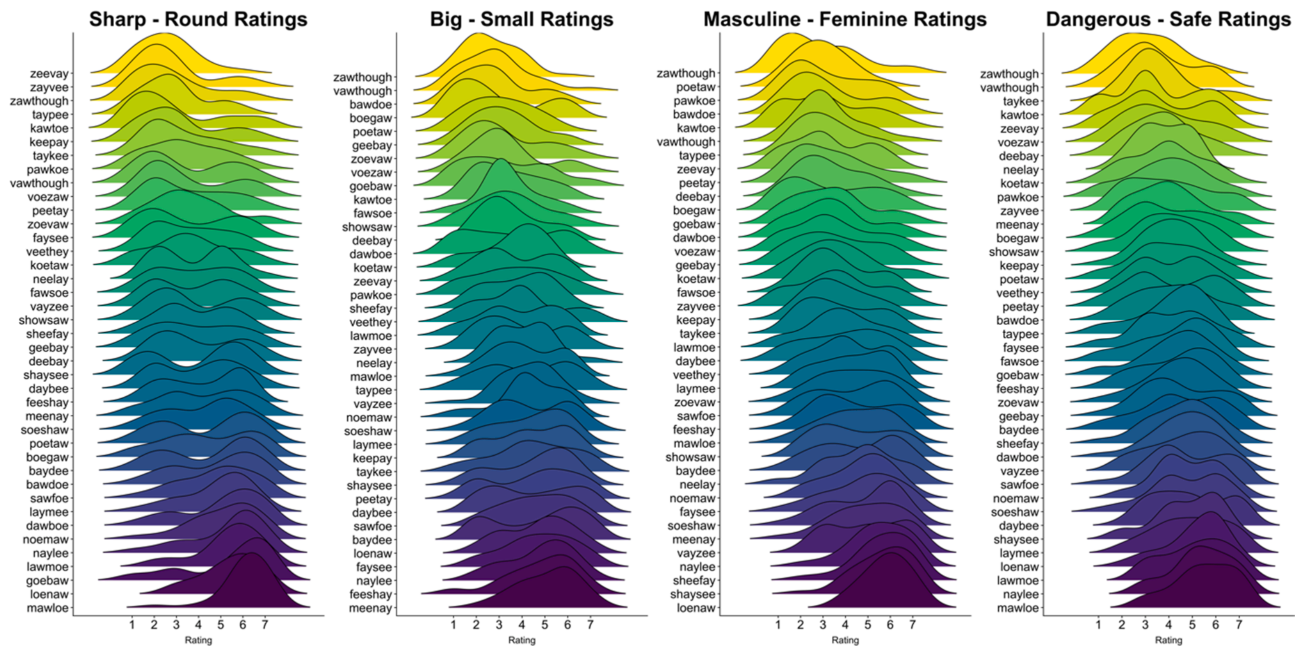

Fig. 4. Distribution of nonword ratings on sharp-round, big-small, masculine-feminine, and dangerous-safe dimensions. We display these dimensions because they are either commonly studied dimensions in sound symbolism (shape, size and gender) or are mentioned in the literature for another reason (e.g., that phonemes may have similarities with animal calls for danger and safety; see Nielsen & Rendall, 2013). Note that nonwords were presented auditorily.

Icon (https://www.flaticon.com/). Unlike nonwords, which were chosen to represent different phoneme categories, there were no such objective categories from which to choose images. However, we wished to include images that evoked a range of associations. Thus, we began with 71 candidate images, and ran a pilot study with a separate group of 58 participants, in which subsets of these images were rated on five semantic differential scales: *pleasant-unpleasant*, *strong–weak*, *fast-slow*, *inhibited-free*, and *ordinary-unique*. We chose 20 shapes that represented a broad range of these dimensions, and also minimized correlations among dimensions. This was to ensure that we included images that evoked a variety of associations. See Fig. 8 for examples. Images were scaled such that their longest side was 11 cm.

Procedure. The procedure was the same as that described for the ratings of List A in Rating Study 1, except that here participants rated images instead of nonwords. Images were presented at the top of the screen.

Results

The data were cleaned in the same manner as the nonword rating data. This led to seven participants being excluded because research assistants noted an issue. We conducted an exploratory factor analysis for the image ratings in the same manner as for the nonwords. As with the nonwords, a four-factor solution was suggested. Because two factors were correlated at $r = .42$, an oblique rotation was used. The fit statistics for our factor analysis suggested a good fit (RMSEA = 0.05, TLI = 0.95). See Table 5 for factor loadings. The first factor is difficult to identify, including elements from valence, potency and activity. However, since the largest loadings correspond to valence, we have tentatively labelled it as such. The next three factors correspond to potency, novelty, and activity. See Table 6 for the correlations among factors.

Nonword-Image fit ratings

Method

Participants. Participants were 67 undergraduate students at the University of Calgary who participated in exchange for course credit. All participants were fluent in English, reported normal or corrected to normal vision and provided informed consent.

Materials. The stimuli were the 20 nonwords from List A, and the 20 rated images.

Procedure. Participants took part in person using the software E Prime. They wore sound attenuating headphones. On each trial, participants first saw a fixation cross for 1000 ms (*ms*). This was followed by a blank screen during which they heard an audio file of the nonword to be rated. This was followed by a blank screen for 250 ms, after which they saw an image. Their task was to rate how well the nonword they had just heard went along with the image, on a scale from 1 (Not at All) to 7 (Very Well). They then saw a blank screen for 500 ms before proceeding to the next trial. Each participant was presented with 200 random trials representing a subset of the potential matches.

Results

The data were cleaned in the same manner as for previous ratings. Six participants were removed for whom research assistants noted an issue, and one participant was removed for indicating that they couldn't focus. Fifty-three trials were removed because participants indicated the nonword was a real word in a language that they spoke.

Our goal was to examine whether nonwords and images with similar scores on the higher order factors were judged as better matches. To that end, we computed average scores for each nonword and image on their extracted four factors. These were then standardized.[3] We computed the Euclidean distance between nonwords and images on each factor. This quantified the similarity between nonwords and images on the four factors, and these values served as predictors in our analyses.

Analyses were done at the trial level of match ratings, and consisted of linear mixed effects models with random subject, nonword and image intercepts. Our predictors of interest were similarity on the four factors. We also included random subject slopes for each predictor. Due to convergence issues, we did not include correlations between slopes and intercepts. See Table 7 for a model summary. Results indicated that nonwords and images that were more similar in terms of the higher

[3] This was done on the set of all 40 nonwords, even though only 20 were included in the present analysis.

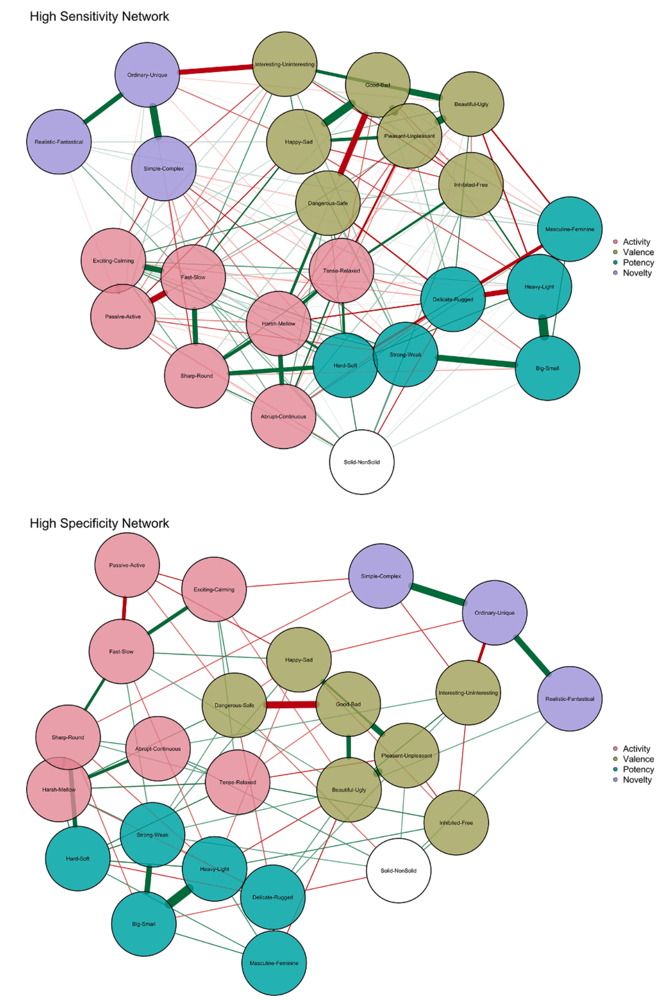

Fig. 5. High sensitivity and high specificity networks describing relationships between dimensions. Note. The top network was computed with high sensitivity to connections (i.e., low Type 2 error) while the bottom network was computed with high specificity (i.e., low Type 1 error). Line thickness corresponds to the size of the partial correlation between dimensions. Line colour corresponds to the direction of the correlation. Green lines correspond to a positive relationship among dimensions (i.e., the first adjectives in a pair of nodes are associated with each other) while red lines correspond to a negative relationship among dimensions (i.e., the first adjective in one node is associated with the second adjective in another node). Node colours correspond to factor loadings from the following exploratory factor analysis (see below). Node placement is the result of an algorithm which aims to have the proximity of nodes correspond to the size of their relationship (see Epskamp et al., 2012). However, this is not an exact correspondence, and should not be over-interpreted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

order factors of novelty (b = -0.07, p = .01), activity (b = -0.05, p = .02), and potency (b = -0.05, p = .045) received higher match ratings. Similarity in terms of valence was not a significant predictor (b = 0.00, p = .93).

In a supplementary analysis we examined whether nonwords and images with similar scores on the 24 *rating dimensions* were judged as better matches. To that end, we first standardized ratings on each scale, separately for nonwords and images, and then computed averages for each nonword and image. We then computed the Euclidean distance between nonwords and images on each dimension. We used these as predictors in a version of the above analysis. To avoid overfitting, we took a stepwise model building approach with backwards selection using the “lmerTest” package. Due to convergence issues, we did not include correlations between random slopes and intercepts, nor a random slope for *hard-soft* distance. This suggested that nonwords and images were judged as better matches if they were given similar ratings on the *masculine-feminine* (b = -0.08, p .008), *ordinary-unique* (b = -0.06, p = .03) and *sharp-round* (b = -0.16, p < .001) scales and if they were given *dissimilar* ratings on the *hard-soft* (b = 0.08, p = .007) scale. See Table 8 for a model summary.

We ran a final analysis that combined the three factor distances that proved useful, with the four scale distances from the previous analysis. Due to convergence issues, we did not include correlations between random slopes and intercepts, nor several random slopes. Of the factors, only novelty remained a significant predictor (b = 0.14, p = .003), while the rating scales of hardness (b = 0.08, p = .02), gender (b = -0.07, p = .03) and shape (b = -0.17, p < .001) all remained significant. See Table 9 for a model summary.

Discussion

We observed the same latent factors underlying the associations of images as the sound symbolic associations of nonwords. Most importantly, the extent to which nonwords and images shared the higher order properties of activity, potency and novelty predicted the rated fit between nonwords and images. However, we found that sharing the dimensions of gender and shape was more predictive of matches. We will return to this in the General Discussion.

Experiment 2b

Another way to examine the effect of higher order dimensions on associations between nonwords and images is to use a forced choice task, presenting shapes that are maximally similar or dissimilar from nonwords on these dimensions. Indeed, forced choice tasks are the most common approach to studying sound symbolism (see Westbury et al., 2018). We adopted that task next.

Method

Participants

Participants were 95 individuals (73 female, 21 male, one gender not recorded; M_{Age} = 23.72, SD_{Age} = 5.23) recruited through the platform Prolific (https://www.prolific.co/). The sample size was determined via an a priori power analysis using the data from Experiment 2a. The ratings collected in that experiment were standardized within participants and ratings between -.5 and .5 were eliminated. The remaining ratings were binarized. We examined how often a participant endorsed nonword image pairings that would be used in this study. This power analysis showed that we would have a power of 100% to detect an effect with 95 participants. In addition, this is a sample size that Trafimow (2018) suggested will have good precision and excellent reliability with a single group. This power analysis was part of the preregistration and can be found here: https://osf.io/3hjz5. Note that 101 participants were tested in total, before we reached a useable sample of 95. The other six participants were excluded for failing an attention check.

Materials

Stimuli consisted of the 40 nonword recordings used in previous experiments. We calculated the mean absolute difference between each nonword and each image on the three factors that were significant predictors of fit ratings (i.e., activity, potency and novelty). Using these values we chose a highly similar and a highly dissimilar image for each nonword. We began with the most similar and dissimilar images, but because a few images were most dissimilar to many nonwords, we set a limit that each image appear on no more than 1/6 of trials. We replaced images with the next most dissimilar (or similar) image until this was true. In addition, we eliminated one image that we judged looked like a pair of lips and thus might bias participants (e.g., on trials including

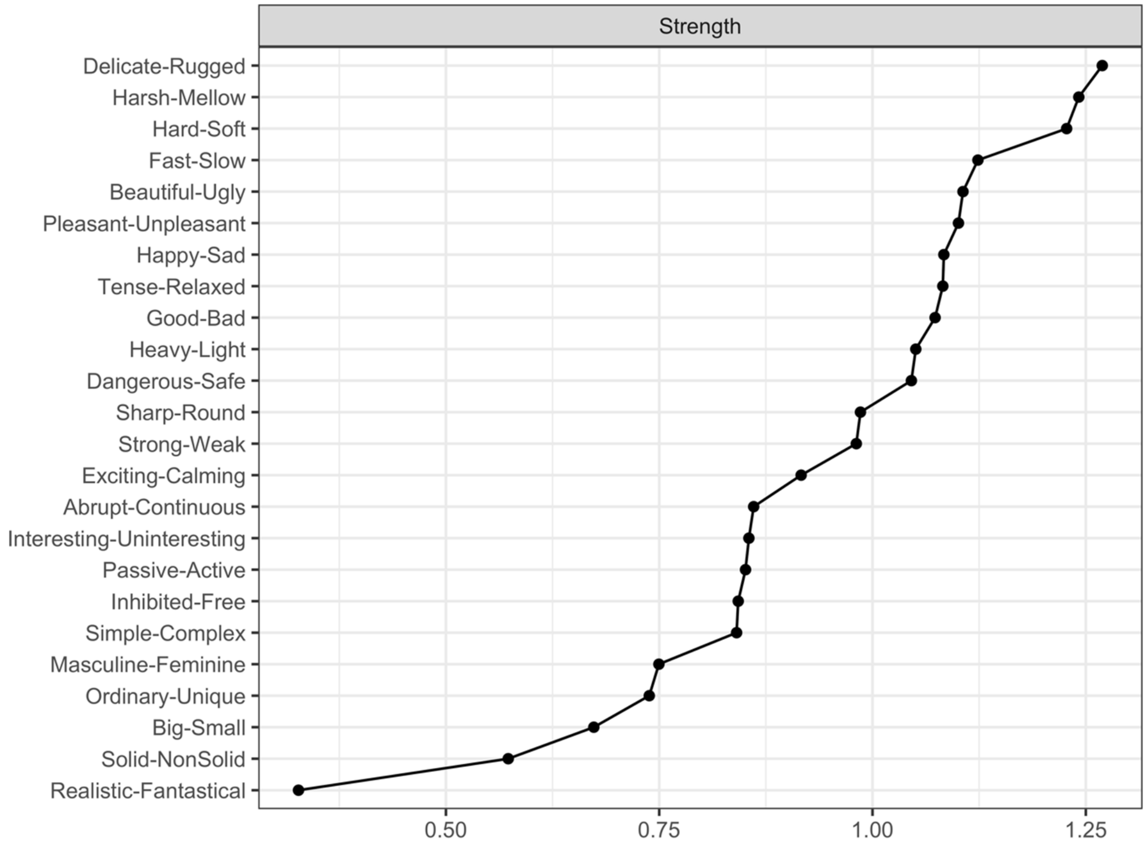

Fig. 6. Centrality of rating dimensions in high sensitivity network. Note. Strength corresponds to the summed value of absolute correlation coefficients for each node in the network.

Table 3
Nonword factor loadings.

Dimension	Activity	Valence	Potency	Novelty
Fast-Slow	−0.75			
Sharp-Round	−0.72			
Exciting-Calming	−0.62			
Harsh-Mellow	−0.57			
Abrupt-Continuous	−0.55			
Tense-Relaxed	−0.55			
Hard-Soft	−0.52			
Passive-Active	0.55			
Pleasant-Unpleasant		−0.76		
Good-Bad		−0.76		
Happy-Sad		−0.72		
Beautiful-Ugly		−0.66		
Interesting-Uninteresting		−0.47		
Dangerous-Safe		0.44		
Inhibited-Free		0.46		
Big-Small			−0.73	
Heavy-Light			−0.7	
Strong-Weak			−0.5	
Masculine-Feminine			−0.42	
Delicate-Rugged			0.44	
Simple-Complex				0.41
Realistic-Fantastical				0.48
Ordinary-Unique				0.65

Note. Only factor loadings > .40 are shown.

bilabials). In addition, we included recordings of the nonwords *maluma* and *takete*, along with a typical round and sharp image for each (see Fig. 1). This was to be able to compare congruent choices for similar/dissimilar images with the classic maluma/takete effect.

Table 4
Nonword factor correlations.

	Activity	Valence	Potency	Novelty
Activity		−0.28	0.50	0.26
Valence	−0.28		−0.45	0.13
Potency	0.50	−0.45		0.05
Novelty	0.26	0.13	0.05	

Procedure

Participants took part online through the survey platform Qualtrics. On each trial they were presented with the recording of a nonword that they could play as many times as they wished, along with two images: a similar and a dissimilar one. Their task was to choose the image that best matched the nonword. Trial order, and the left/right placement of image pairs, was randomized across participants. Participants also answered a debriefing questionnaire which asked them if any of the nonwords were real words in a language they spoke.

Results

Six trials were removed because participants indicated the nonword was a real word in a language that they spoke, or was another real word that they knew (e.g., a singer named Maluma). We ran an analysis at the trial level, excluding *maluma* and *takete* trials. This consisted of a logistic mixed effects model, with random subject and nonword intercepts. The dependent variable was whether a participant chose the highly similar shape on a given trial. This model had a significant intercept ($b = 0.60$, $p < .001$), indicating that participants were 1.83 times more likely to

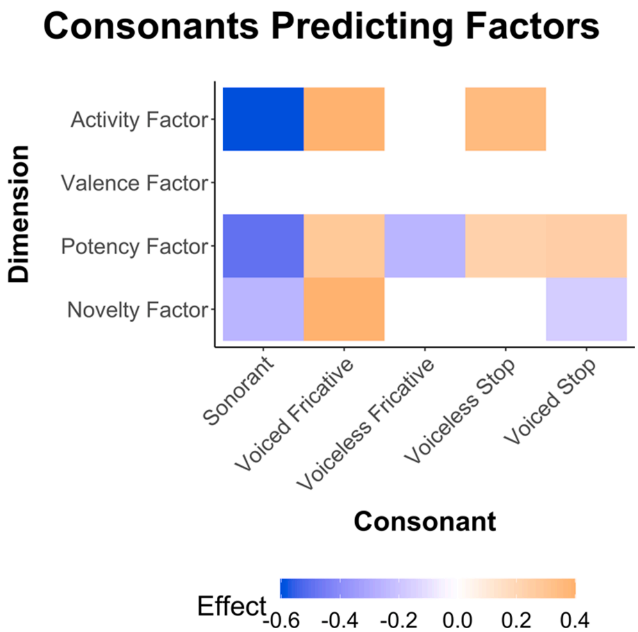

Fig. 7. Results of consonant categories predicting factor scores. In particular, the differences between the estimated marginal means for each consonant category and overall mean are shown. Note. Warm (cool) colours indicate that a consonant category was associated with the high (low) end of a given factor, compared to the mean across all nonwords.

choose the similar as opposed to the dissimilar image. Participants chose the similar image on 64.43% of trials. For comparison, a simple logistic regression[4] conducted on *maluma* and *takete* trials also found a significant intercept ($b = 1.74$, $p < .001$) and that participants were 5.71 times more likely to choose the congruent shape. Participants chose the congruent image on 85.11% of *maluma* and *takete* trials.

In a supplementary analysis, we examined if the difference in similarity between images on a given trial affected participants' likelihood of choosing the similar image. This analysis consisted of a logistic mixed effects regression with the standardized difference in similarity between each image and the nonword as the predictor of interest. It also included a random subject slope for difference, and random subject and item intercepts. Due to convergence issues, we did not include correlations between slopes and intercepts. Note that because we chose highly dissimilar pairs, there was not a great deal of variance in similarity difference. Difference was a marginally significant predictor ($b = 0.09$, $p = .051$), with participants 1.09 times more likely to choose the similar image for every one point increase in standardized similarity difference.

Table 5
Image factor loadings.

Dimension	Valence	Potency	Novelty	Activity
Pleasant-Unpleasant	−0.86			
Good-Bad	−0.84			
Beautiful-Ugly	−0.82			
Happy-Sad	−0.73			
Delicate-Rugged	−0.51			
Structured-Disorganized	−0.43	−0.44		
Hard-Soft	0.47	−0.42		
Masculine-Feminine	0.5			
Abrupt-Continuous	0.52			
Tense-Relaxed	0.62			
Harsh-Mellow	0.62			
Dangerous-Safe	0.67			
Heavy-Light		−0.67		
Strong-Weak		−0.64		
Solid-NonSolid		−0.60		
Big-Small		−0.58		
Interesting-Uninteresting	0.47		−0.54	
Realistic-Fantastical			0.57	
Simple-Complex			0.69	
Ordinary-Unique			0.77	
Fast-Slow				−0.70
Exciting-Calming				−0.52
Passive-Active				0.53

Note. Only factor loadings > .40 are shown.

Table 6
Image factor correlations.

	Valence	Potency	Novelty	Activity
Valence		−0.42	0.10	−0.20
Potency	−0.42		−0.12	0.32
Novelty	0.10	−0.12		0.38
Activity	−0.20	0.32	0.38	

Discussion

When given the choice between an image that was highly similar and highly dissimilar to a given nonword on these factors, participants tended to choose the highly similar image as being the better match. However, this effect was smaller than the one typically observed in studies on the maluma/takete effect.

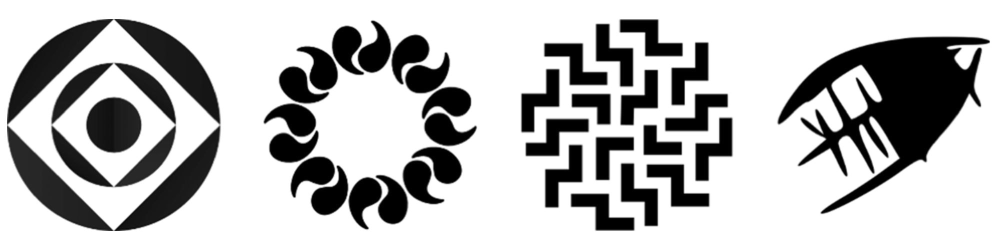

Fig. 8. Example abstract shape stimuli. Note. From left to right: Logo made by "RoundIcons" from http://www.flaticon.com/; Saitama Japan Flag Symbol made by "FreePik" from https://www.flaticon.com/; "abstract labyrinth" by Alice Noir, from the Noun Project; "petroglyph" by Solar Map Project, from the Noun Project.

[4] We omitted random subject intercepts in this study because there were only two observations per subject. Including them leads to a coefficient of 5.55 implying participants were 257.24 times more likely to select the congruent shape.

Table 7
Results from linear mixed effects model predicting nonword-image fit using factor distance.

Fixed Effect	B	SE	t	p
Intercept	3.50	0.11	31.95	<.001***
Activity Factor Distance	−0.05	0.02	−2.34	.02*
Valence Factor Distance	0.00	0.03	0.09	.93
Potency Factor Distance	−0.04	0.02	−2.02	.045*
Novelty Factor Distance	−0.07	0.03	−2.52	.01*

Random Effect	s^2
Subject Intercept	0.47
Subject Activity Slope	0.01
Subject Valence Slope	0.01
Subject Potency Slope	0.00
Subject Novelty Slope	0.01
Image Intercept	0.06
Nonword Intercept	0.02

Table 8
Results from linear mixed effects model predicting nonword-image fit using rating scale distances.

Fixed Effect	B	SE	t	p
Intercept	3.50	0.11	30.59	<.001***
Hard-Soft Distance	0.08	0.03	2.72	.007**
Masculine-Feminine Distance	−0.08	0.03	−2.69	.008**
Ordinary-Unique Distance	−0.06	0.03	−2.26	.03*
Sharp-Round Distance	−0.16	0.03	−5.04	<.001***

Random Effect	s^2
Subject Intercept	0.47
Subject Masculine-Feminine Slope	0.01
Subject Ordinary-Unique Slope	0.01
Subject Sharp-Round Slope	0.02
Image Intercept	0.08
Nonword Intercept	0.02

Table 9
Results from linear mixed effects model predicting nonword-image fit using both higher order dimensions and rating scale distances.

Fixed Effect	B	SE	t	p
Intercept	3.50	0.11	30.99	< .001***
Hard-Soft Distance	0.08	0.03	2.41	.016*
Masculine-Feminine Distance	−0.07	0.03	−2.20	.029*
Ordinary-Unique Distance	0.04	0.05	0.89	.37
Sharp-Round Distance	−0.17	0.04	−4.73	< .001***
Activity Factor Distance	0.04	0.03	1.39	.17
Potency Factor Distance	−0.01	0.03	−0.28	.78
Novelty Factor Distance	−0.14	0.05	−3.01	.003**

Random Effect	s^2
Subject Intercept	0.47
Subject Masculine-Feminine Slope	0.01
Subject Ordinary-Unique Slope	0.01
Subject Sharp-Round Slope	0.02
Subject Activity Slope	0.01
Image Intercept	0.08
Nonword Intercept	0.02

General Discussion

The main goal of the present paper was to examine whether there are higher order semantic factors at play in sound symbolism. We explored this by measuring the associations between a set of nonwords and 25 different semantic dimensions in Experiment 1. We then also examined whether these factors play a role in the association of nonwords and abstract images in Experiments 2a and 2b. This allowed us to address several research questions. We will next summarize our findings with regards to each of these questions and then discuss their broader theoretical implications.

Associations of different phoneme categories

We found that all categories of phonemes had associations with some semantic dimensions. Several of these replicated well-known effects. For example, sonorants, voiced stops and back vowels were associated with roundness, while voiceless stops, voiced fricatives and front vowels were associated with sharpness (i.e., the maluma/takete effect). These associations have been demonstrated previously (e.g., McCormick et al., 2015), but it is noteworthy that they emerged here even when tested in the midst of 25 different ratings scales. An interesting point is that sonorants' association with roundness was the largest of any of the associations observed. This could be a reason that shape sound symbolism has emerged as the prototypical example of sound symbolism. These results also support the proposal that neither voicing nor manner of articulation alone explain the maluma/takete effect (e.g., McCormick et al., 2015; Monaghan & Fletcher, 2019). For instance, some voiced consonants were associated with roundness (i.e., stops) while others were associated with sharpness (i.e., fricatives). It seems that a phoneme's association will depend on its specific combination of features (see Monaghan & Fletcher, 2019). There may be some other property that emerges from the combination of features that causes a phoneme's association with shape. For instance, voiceless stops and voiced fricatives may both be associated with sharpness due to their discontinuous/strident sounds (see McCormick et al., 2015).

With also found several associations between consonants and size: sonorants and voiceless fricatives were associated with smallnes, while voiced fricatives and voiced stops were associated with largeness. In addition, we replicated the frequently reported association between front vowels and smallness, and back vowels and largeness (Newman, 1933; Sapir, 1929; i.e., the mil/mal effect). Another notable dimension was gender. Sonorants and voiceless fricatives showed an association with feminineness, while both voiceless and voiced stops showed an association with maleness. Interestingly, this contrasts with a finding by Slepian and Galinsky (2016) that American male (female) names were more likely to begin with a voiced (voiceless) consonant. It also contrasts with a study of brand names, which found that a brand name was judged as more feminine if it contained a voiceless vs. a voiced stop (Klink, 2000). However, the pattern observed in the present study is consistent with the finding in Klink (2000) that product names containing voiceless vs. voiced fricatives were judged as more feminine.

Some dimensions were not associated with any particular kinds of phonemes. In terms of consonants, these were each of the valence dimensions. This may suggest that there is no obvious mapping between the features of consonants and valence.[5] That is, there may not be an obvious perceptuomotor analogy that would allow an association between consonant phonemes and valence. An exception to this, with regards to vowels, is the proposed overlap in facial movements used to express emotion and articulate language (i.e., smiling when articulating /i/; Rummer et al., 2014). In fact, we did observe an association between front/back vowels and the specific valence dimension *happy-sad*. This specific finding (in the absence of and associations with *good-bad* or *pleasant-unpleasant*) may speak to emotional facial expressions as a mechanism for valence sound symbolism in vowels.

[5] In a supplementary analysis we examined the possibility that nonwords might be associated with these dimensions based on the direction of their articulation (i.e., inward as in *peetay* vs outward as in *taypee*; see Topolinski et al., 2014). Direction of articulation was not a significant predictor for any of the valence dimensions, whether including all nonwords that could be coded for direction or only those including stops (p's > .38).

In general, sonorants, voiced fricatives and voiceless stops were the categories of phonemes with the most associations. In addition, sonorants and voiced fricatives/voiceless stops appeared at the opposite ends of dimensions. This may suggest that these phonemes are maximally dissimilar with regards to whatever phoneme quality is being associated with semantic dimensions. Based on these categories of phonemes we might speculate that the relevant quality is phonemes' resonance and/or continuity of sound. As in the case of the *sharp-round* dimension, voicing alone doesn't seem to explain these patterns.

Higher order factors in phonemes' associations

We found evidence that there are higher order semantic factors among the sound symbolic associations we measured. Namely, sound symbolic associations clustered into four factors that we have termed activity, valence, potency and novelty.[6] These results suggest that sound symbolic associations can be grouped according to the same higher order factors that have been shown to exist for word meaning. This is yet another demonstration that the factors discovered by Osgood et al. generalize beyond word stimuli. We also found support for a fourth factor (i.e., novelty) that has sometimes been found in the past (e.g., the familiarity factor in Bentler & Lavoie, 1972). Other work has suggested intensity (Tzeng et al., 2017) or magnitude (see Spence, 2011) as potentially relevant higher order dimensions in crossmodal matching. However, these do not seem to be consistent with the present results.

Associations between phoneme categories and higher order factors

Importantly, we found that certain phoneme categories were associated with these higher order factors. High (low) activity was associated with voiced fricatives, voiceless stops and front vowels (sonorants and back vowels); high (low) potency was associated with voiced fricatives, stops and back vowels (sonorants, voiceless fricatives and front vowels). This may suggest that there are (at least) two distinct clusters of sound symbolism effects (i.e., those related to activity and those related to potency). More importantly, this suggests that the various sound symbolic associations of a phoneme are not entirely distinct phenomena. One interpretation of these data is that various sound symbolic associations are the result of a few basic associations between phonemes and higher order semantic factors. In other words, phonemes may not have distinct associations with, for example, speed, shape, excitement. Instead, these specific associations may be in part determined by phonemes' associations with the higher order factor of activity.

In the Introduction, we outlined two mechanisms that have been proposed for sound symbolism: statistical co-occurrence and shared properties. The present results appear to be consistent with the mechanism of shared properties, namely the shared higher order properties of activity and potency. That is, what unites voiceless stops and sharpness, for example, may be the higher order factor of activity. This is consistent with recent findings by Aryani et al. (2020; see also Aryani et al., 2018) suggesting that arousal may explain mappings in the maluma/takete effect. Of course, higher order factors cannot explain all of sound symbolism. For instance, the communality for the *sharp-round* dimension was 0.56, suggesting that 56% of variance in sharp-round is explained by higher order factor scores (see Table S2 in Supplementary Material for each dimension's communality). Thus, while associations between phonemes and higher order semantic factors may explain part of sound symbolic associations, other mechanisms will still be at play. For example, the visual similarity between round vowels' articulation and round shapes is not likely to be captured by the activity factor. It is not possible to rule out any potential mechanisms based on these data.

A key question is how phonemes become associated with higher order semantic factors. At this point we can only offer speculation. It is possible that phonemes are connected to these higher order factors via perceptuomotor analogies. For instance, voiceless stops' association with high activity could reflect an analogy between their abrupt onset and the energetic potential defining high activity. Another possibility is that phonemes tend to co-occur with various stimuli sharing a given factor in the real world. For example, it could be that various events or objects in the world representing high activity make sounds that are more similar to voiceless stops than sonorants. In this case, an association between voiceless stops and high activity would be due to an internalization of these statistical regularities.

A potentially informative observation is that the most central dimensions in our network analysis were *delicate-rugged, harsh-mellow*, and *hard-soft.* These dimensions showed the most association with others in the network. Interestingly, each of these dimensions can be used to describe the sensorimotor properties of phonemes. An intriguing possibility is that these dimensions are primary to sound symbolism. That is, for example, a phoneme's articulation might feel particularly *hard*. This may then associate that phoneme to the various dimensions associated with *hard-soft.*

Higher order factors and the fit between nonwords and visual stimuli

Although not the main focus of this paper, in Experiments 2a and 2b, we also explored whether these higher order dimensions could explain associations between phonemes and perceptual stimuli. A similar explanation was put forth for associations between phonemes and different tastes (see Gallace et al., 2011). In Experiment 2a, we observed that when phonemes and abstract shapes were associated with the same higher order semantic factors, they were rated as being a good fit for one another. However, similarity on the individual dimensions of gender and shape proved to be better predictors of match ratings than higher order dimensions.[7] This would suggest that associations between nonwords and perceptual stimuli are not primarily determined at the level of latent higher order dimensions. One might speculate that such dimensions will play more of a role when pairings require a greater amount of abstraction (e.g., between nonwords and personality traits; Sidhu et al., 2019).

It is not entirely surprising that the dimension of shape is a good predictor of nonword-abstract shape pairings. Presumably when nonwords were rated on the dimension of *sharp-round*, participants were rating precisely this: fit between nonwords and visual roundness/ sharpness. Thus, there is a danger of circularity here, in a sense using shape sound symbolism ratings to explain shape sound symbolism ratings. However, the finding that gender can explain the fit between nonwords and shapes is more interesting. Gender may be functioning as a higher order dimension here, though at a lower level of abstraction than the factors we extracted. Of course, this is pure speculation and should be investigated further in future research.

In Experiment 2b, when given the choice between shapes that were highly similar vs. dissimilar to a given nonword on higher order semantic factors, participants tended to choose the highly similar one. However, this effect was not as large as the classic maluma/takete effect (64% vs. 85%, respectively). This may be because the classic maluma/ takete shapes offer a very salient shape contrast while the shapes that we used here were more complex and differed on a variety of visual dimensions. While *in aggregate* the shapes that we used here differed on higher order factors, this may be less salient when a variety of lower order contrasts are present. This is consistent with the dimension *sharp-round* playing a major role in decisions (leading to larger effects when there is a salient contrast on this dimension).

[6] Note that *solid-nonsolid* did not load onto any factors.

[7] As to why *dissimilarity* on the dimension of hardness was a predictor of match ratings, we can only offer the very tenuous speculation that this dimension could manifest differently for nonwords and images.

Limitations

In future research it will be important to explore whether these results generalize beyond the nonword stimuli that we used. In particular, because we used a small set of 40 nonwords, each containing consonants and vowels from only one category, we may have accentuated effects of sound symbolism (see Westbury et al., 2018). In addition, because nonwords were categorically distinct, participants may have been encouraged to compare and contrast their phonology, which could also have accentuated effects. A future study might take a big data approach (e.g., Westbury et al., 2018) and examine a large number of less constrained nonwords. This would also be an ideal way to examine potential interactions between consonant and vowel effects.

In addition, the factors that we extracted may have in part been due to the 25 dimensions that we chose. We made an effort to choose dimensions that broadly sampled from different kinds of meanings. Nevertheless, when choosing scales for this experiment, we selected three each to represent the factors of valence, activity, and potency. In contrast, we only selected one scale to represent other factors (e.g., complexity). This may have biased the factor analysis towards extracting the factors that were over-represented among the scales we included.

It is also important to note that the labelling of factors is an inherently subjective process. Therefore, there may have been other possible labels for the factors that we extracted. Here we erred on the side of using established labels for our first three factors (i.e., activity, valence and potency) though it is possible that other labels would have been more accurate. In addition, we used a semantic differential approach because we judged that to be the best way to characterize sound symbolic associations. There are certainly other ways of generating semantic ratings (e.g., feature listing, generating associates). It is possible that another experimental approach could have led to different results.

Conclusion

Here we have shown that sound symbolic associations group according to the higher order factors of activity, potency and novelty. Importantly, categories of phonemes have detectable associations with these higher order factors. Research continues to identify new sound symbolic associations. The time has come for the field to consider how these various associations fit together. Doing so has the potential to contribute to our understanding of the basic mechanisms at work in sound symbolism.

CRediT authorship contribution statement

David M. Sidhu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Gabriella Vigliocco:** Conceptualization, Writing – review & editing, Supervision. **Penny M. Pexman:** Conceptualization, Methodology, Resources, Writing – review & editing, Project administration, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and analysis code are available at: https://osf.io/gruqs/.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to P.M. P. (RGPIN/03860-2018), and a Graduate Scholarship to D.M.S. (CGSD3 476111 2015). It was also supported by Alberta Innovates Health Solutions (AIHS) through a Graduate Scholarship to D.M.S. (201500125-1 CA# 3874). Gabriella Vigliocco was supported by a European Research Council Advanced Grant (ECOLANG, 743035) and Royal Society Wolfson Research Merit Award (WRM\R3\170016). The authors would like to thank Summer Abdalla, Ahmad Alawami, Natalia Czarnecki, and Kelsey Friesen for their help with data collection; and Jacqueline Beatch and Ellen Cribbs for recording the nonwords.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2022.104323.

References

Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why 'piss' is ruder than 'pee'? The role of sound in affective meaning making. *PloS one, 13*.
Aryani, A., Isbilen, E. S., & Christiansen, M. H. (2020). Affective arousal links sound to meaning. *Psychological Science, 31*, 978–986.
Auracher, J. (2017). Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance. *PLOS ONE, 12*. https://doi.org/10.1371/journal.pone.0187196
Bartoń, K. (2020). MuMIn: Multi-Model Inference. *R package version, 1*(43), 17.
Bentler, P. M., & LaVoie, A. L. (1972). An extension of semantic space. *Journal of Verbal Learning and Verbal Behavior, 11*, 174–182.
Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences, 136*, 68–78.
Cuskley, C. (2013). Mappings between linguistic sound and motion. *Public Journal of Semiotics, 5*, 39–62.
Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., … Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. In *Philosophical Transactions of the Royal Society B* (p. 377).
Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*, 1–18.
Epskamp, S., Fried, E. I., & Epskamp, M. S. (2015). Package 'bootnet'.
Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*, 617–634.
Estes, Z., & Adelman, J. S. (2008). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion, 8*, 441–444.
Fang, S., Murumatsu, K., & Matsui, T. (2015). Experimental study of aesthetic evaluation to multi-color stimuli using semantic differential method. *Transactions of Japan Society of Kansei Engineering, 14*, 37–47.
Fort, M., Lammertink, I., Peperkamp, S., Guevara-Rukoz, A., Fikkert, P., & Tsuji, S. (2018). Symbouki: A meta-analysis on the emergence of sound symbolism in early language acquisition. *Developmental Science, 21*, Article e12659.
French, P. L. (1977). Towards an explanation of phonetic symbolism. *Word, 28*, 305–322. https://doi.org/10.1080/00437956.1977.11435647
Gallace, A., Boschin, E., & Spence, C. (2011). On the taste of "Bouba" and "Kiki": An exploration of word-food associations in neurologically normal participants. *Cognitive Neuroscience, 2*, 34–46. https://doi.org/10.1080/17588928.2010.516820
Greenberg, J. H., & Jenkins, J. J. (1966). Studies in the psychological correlates of the sound system of American English. *Word, 22*, 207–242.
Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*, 1744–1756.
Isvoranu, A. M., & Epskamp, S. (2021). Which estimation method to choose in network psychometrics? Deriving guidelines for applied researchers. *PsyArXiv*. doi:10.31234/osf.io/mbycn.
Kim, H. W., Nam, H., & Kim, C. Y. (2018). [i] is lighter and more greenish than [o]: Intrinsic association between vowel sounds and colors. *Multisensory Research, 31*, 419–437.
Klink, R. R. (2000). Creating brand names with meaning: The use of sound symbolism. *Marketing Letters, 11*, 5–20.
Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports, 7*, 1–11.
Köhler, W. (1929). *Gestalt Psychology*. New York, USA: Liveright.
Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition, 112*, 473–481.
Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143*, 1065–1081.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.6.2-1.

Malhotra, N. K. (1981). A scale to measure self-concepts, person concepts, and product concepts. *Journal of Marketing Research, 18*, 456–464.
McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). In *Sound to meaning mappings in the bouba–kiki effect* (pp. 1565–1570). Austin, TX: Cognitive Science Society.
McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*, 99–130.
Miron, M. S. (1961). A crosslinguistic investigation of phonetic symbolism. *The Journal of Abnormal and Social Psychology, 62*, 623–630. https://doi.org/10.1037/h0045212
Monaghan, P., & Fletcher, M. (2019). Do sound symbolism effects for written words relate to individual phonemes or to phoneme features? *Language and Cognition, 11*, 235–255.
Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*.
Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution, 4*, 133–142.
Nielsen, A. K. S., & 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, 65*, 115–124. https://doi.org/10.1037/a0022268
Nielsen, A. K., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology, 67*, 153–163.
Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology, 45*, 53–75.
Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*, 1–8.
Osgood, C. E., Suci, S. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning.* Urbana, IL: Univ. of Illinois Press.
Pejovic, J., & Molnar, M. (2016). The Development of Spontaneous Sound-shape Matching in Monolingual and Bilingual Infants During the First Year. *Developmental Psychology*. https://doi.org/10.1037/dev0000237
Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.9.
Rummer, R., Schweppe, J., Schlegelmilch, R., & Grice, M. (2014). Mood is linked to vowel type: The role of articulatory movements. *Emotion, 14*, 246–250.
Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225–239.
Sidhu, D. M., Deschamps, K., Bourdage, J. S., & Pexman, P. M. (2019). Does the name say it all? Investigating phoneme-personality sound symbolism in first names. *Journal of Experimental Psychology: General, 148*, 1595–1614.
Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? Sound symbolism and gender in first names. *PloS ONE, 10*. https://doi.org/10.1371/journal.pone.0126809
Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review, 25*, 1619–1643.
Slepian, M. L., & Galinsky, A. D. (2016). The voiced pronunciation of initial phonemes predicts the gender of names. *Journal of Personality and Social Psychology, 110*, 509–527.
Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995. https://doi.org/10.3758/s13414-010-0073-7
Styles, S. J., & Gawne, L. (2017). When does maluma/takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception, 8*. https://doi.org/10.1177/2041669517724807
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
Topolinski, S., Maschmann, I. T., Pecher, D., & Winkielman, P. (2014). Oral approach–avoidance: Affective consequences of muscular articulation dynamics. *Journal of Personality and Social Psychology, 106*, 885–896.
Trofimova, I. (2014). Observer bias: An interaction of temperament traits with biases in the semantic perception of lexical material. *PloS one, 9*.
Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology, 31*, 1188–1214.
Tzeng, C. Y., Nygaard, L. C., & Namy, L. L. (2017). The Specificity of Sound Symbolic Correspondences in Spoken Language. *Cognitive Science, 41*, 2191–2220.
Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). The meanings of object and action words. *Cognitive Psychology, 48*, 422–488.
Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing? *Cognition & Emotion, 28*, 737–746.
Walker, L., Walker, P., & Francis, B. (2012). A common scheme for cross-sensory correspondences across stimulus domains. *Perception, 41*, 1186–1192.
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191–1207.
Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language, 99*, 122–150.
Wickens, D. D., & Lindberg, M. A. (1975). A test of four proposed new dimensions of semantic space. *Bulletin of the Psychonomic Society, 6*, 381–382.
Williams, D. R., & Mulder, J. (2019). "BGGM: Bayesian Gaussian Graphical Models in R." *PsyArXiv*. R package version 2.0.4.
Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review, 14*, 1218–1225.