# Modelling Form-Meaning Systematicity with Linguistic and Visual Features

**Arie Soeteman**
University of Amsterdam
awsoeteman@hotmail.com

**Dario Gutierrez**
IBM Research
elkin.gutierrez@ibm.com

**Elia Bruni**
University of Amsterdam
e.bruni@uva.nl

**Ekaterina Shutova**
University of Amsterdam
e.shutova@uva.nl

## Abstract

Several studies in linguistics and natural language processing (NLP) pointed out systematic correspondences between word form and meaning in language. A prominent example of such systematicity is iconicity, which occurs when the form of a word is motivated by some perceptual (e.g. visual) aspect of its referent. However, the existing data-driven approaches to form-meaning systematicity modelled word meanings relying on information extracted from textual data alone. In this paper, we investigate to what extent our visual experience explains some of the form-meaning systematicity found in language. We construct word meaning representations from linguistic as well as visual data and analyze the structure and significance of form-meaning systematicity found in English using these models. Our findings corroborate the existence of form-meaning systematicity and show that this systematicity is concentrated in localized clusters. Furthermore, applying a multimodal approach allows us to identify new patterns of systematicity that have not been previously identified with the text-based models.

## 1 Introduction

Linguistic arbitrariness refers to the notion that there are no structural ties between words and their meanings. The concept of arbitrariness has been at the foundation of numerous linguistic theories ever since its introduction in the field of linguistics by Ferdinand de Saussure in 1916 (De Saussure and Baskin 1916). In *The Origin of Speech*, for example, Charles Hockett (1960) argued that arbitrariness in "the ties between meaningful message-elements and their meaning" is a universal feature of human language, and a necessary condition for extensive and flexible communication (Hockett and Hockett 1960). Gasser (2004) introduced a formalized notion of arbitrariness as the absence of iconicity (i.e. direct similarity relations between word form and meaning). He argued that iconicity can facilitate early language acquisition since it reduces the information required to learn form-meaning mappings. However, when a language becomes more extensive, iconicity limits the available space for new words, i.e. the number of forms that can be related to a meaning. This limitation of space causes overlap between the forms used for different meanings, and therefore leads to ambiguity. This ambiguity hinders language acquisition by obscuring form-meaning relations. Based on formal simulations Gasser therefore reached the same conclusion as Hockett: arbitrariness is a fundamental property of extensive languages (Gasser 2004).

However, linguists have long noted exceptions to arbitrariness. For instance, vowels with high acoustic frequency tend to be associated with smallness, while vowels with low acoustic frequency tend to be associated with largeness (Ohala 1984). Onomatopoeic words (e.g. 'bark', 'cling', 'clang', 'slurp') directly echo the sound of their referent. Phonaesthemes are another counterexample to arbitrariness in language. They represent phonetic clusters that occur in words with related meanings. For instance, numerous English words that start with 'sn-' are related to the nose and mouth (e.g. 'snore', 'snorkel', 'sniff', 'snout', 'snot'). Otis (2008) statistically confirmed the existence of 27 of such phonaesthemes in English, while research into the psychological reality of phonaesthemes found that native English speakers perceive as many as 46 (Hutchins 1999). While these studies point towards localized clusters of non-arbitrariness, they do not address the role of arbitrariness in a language as a whole. Richard Shillcock (2001) first used distributional semantics to analyze the significance of arbitrariness in the English language, by correlating the distance between distributional semantic vectors with orthographic distance. This approach has been further developed by Monaghan et al. (2014) and Gutierrez et al. (2016). All three studies found a small but statistically significant correlation between form and meaning. However, Monaghan et al. (2014) found systematicity to be diffusely distributed across language, while Gutierrez et al. (2016) found systematicity to be concentrated in localized phonological clusters.

The semantic vectors used to represent word meaning in these three studies are based on word usage in textual corpora. However, research into grounded cognition suggests that simulations of visual, auditory and other sensorimotor stimuli play a substantial role in language comprehension (Barsalou 2008).

For instance, Gernsbacher et. Al (1990) found that individuals abilities for comprehending events visually versus verbally were highly correlated. More recently, research into our cognitive capacities for text representation (Zwaan and Madden 2005; Fincher-Kiefer and D'Agostino 2004) has supported the theory that readers construct visual simulations to represent text. Furthermore, Fincher-Kiefer and D'Agostino found that language comprehension is influenced by whether interfering or noninterfering visual information is maintained in working memory while reading. These findings have motivated the development of multimodal semantic models (Bruni, Tran, and Baroni 2014; Kiela and Bottou 2014), in which vector representations of words are not merely based on their usage in textual corpora, but also on visual and auditory data. Previous research on arbitrariness (Ohala 1984; Gutiérrez, Levy, and Bergen 2016; Gasser 2004) has also identified systematic relations between word form and visible properties of the corresponding referents in their qualitative analyses. Furthermore, neurological and behavioral research has shown that visual information plays a role in semantically representating words in human cognition (Barsalou 2008; Zwaan and Madden 2005). This suggests that by incorporating visual information into semantic representations, more insight can be gained into the systematic relations between form and meaning that exist in language.

In this paper, we propose the first multimodal approach to linguistic arbitrariness, extending the method of Gutierrez et al (2016) by incorporating visual features into semantic representations. We use a text-based model trained using Skipgram with negative sampling (Mikolov et al. 2013), and construct an image-based model using a convolutional neural network (CNN). We first analyze form-meaning systematicity using the text-based and image-based models separately. We then combine the two models into three seperate multimodal models using scoring-level fusion, dimensionality reduction and concatenation, and multimodal fusion using a neural network. Experimenting on the English lexicon, we find that the inclusion of visual features allows us to identify more form-meaning systematicity than when using a text-based model alone. Furthermore, the multimodal model discovers different relations between form and meaning than the two monomodal models, and identifies multiple novel phonaesthemes.

## 2 Related work

### 2.1 Data-driven approaches to form-meaning systematicity

Shillcock et al. (2001) were the first to quantify the significance of form-meaning systematicity in a lexicon by computing the correlation between phonological and semantic distances of English monosyllabic words. They found a small but statistically significant correlation. Additionally, they calculated this same correlation after omitting each individual word from the dataset to determine the systematicity of that word. The extent to which the correlation decreased after omitting the word was taken as a measure of the word's systematicity. They found that many of the words deemed systematic by this method tend to be communicatively important.

In a subsequent study, Monaghan et al. (2014) made use of the Mantel test for pairwise distances (Mantel 1967) to compute the correlation between form and meaning. Similar to Shillcock et al. (2001), they found a small (r = 0.016) but statistically significant correlation. Furthermore, they found comparable correlations using different distance metrics, emphasizing the robustness of the results. Using the same methods for evaluating the systematicity of individual words as Shillcock et al., they found that systematicity is not concentrated in localized clusters, but is a property of language as a whole.

The aforementioned study of Gutierrez et al. (2016) expanded on the methods of Shillcock et al (2001) and Monaghan et al. (2014). Instead of using phonological edit distances, they analyzed edit distances between orthographic strings in English monomorphemic words. They optimized weights to represent the difference in semantic relevance between string edits. Furthermore, they proposed a new method for evaluating the systematicity of individual words: assessing the extent to which their semantic vectors can be predicted based on their strings. Edit weight optimization increased the resulting correlation between word form and meaning (from r = 0.0194 to r = 0.0464). Furthermore, using their new evaluation method for the systematicity of individual words Gutierrez et al. found numerous localized clusters in which systematicity is concentrated. This contradicted the conclusion of Monaghan et al. (2014) that systematicity is diffusely distributed across language.

A different approach was developed by Liu et. al (2018), who implemented linear regression to directly predict semantic vectors from binary feature vectors that encode the submorphemes occuring in a word. They used sparse regularization to select semantically relevant features, resulting in a list of 30 phonaesthemes. In addition, they introduced a two-step model in which the components of semantic vectors that are predictable from morpheme-level information are removed, enabling the use of polymorphemic words in the lexicon. However, while this is an effective method for identifying form-meaning relations, it does not provide a measure for the significance of form-meaning systematicity within the lexicon as a whole.

### 2.2 Multimodal semantic models

All of the above corpus-wide studies on form-meaning systematicity (Gutiérrez, Levy, and Bergen 2016; Monaghan et al. 2014; Shillcock et al. 2001) were performed using distributional semantic models in which word meanings were approximated by vectors that represent their usage in textual corpora. This approach is based on the distributional hypothesis, which states that words that occur in similar contexts are semantically similar (Harris 1970; Wittgenstein 1953). However, recent research has shown that multimodal semantic models that learn from both linguistic and visual data outperform the purely text-based models on tasks such as word categorization (Bruni, Tran, and Baroni 2014; Kiela and Bottou 2014), lexical entailment (Kiela et al. 2015), modelling compositionality (Roller and Schulte im Walde 2013) and metaphor identification (Shutova, Kiela, and Maillard 2016). The currently best-performing method for constructing image-based word vectors extracts visual features from images using convolutional neural networks (Kiela 2016). In order to use a CNN for visual feature extraction, it is pre-trained on an image-classification or -regression task on a large labeled dataset such as ImageNet (Russakovsky et al. 2015). Subsequently, it is applied to images that represent the words for which visual vectors are to be extracted. The layer preceding the final softmax classification layer is then extracted and stored as a vector representation of the image. A visual vector for a given word is constructed by combining the vectors of all relevant images (Kiela and Bottou 2014).

Several methods have been developed for combining linguistic and visual information into a multimodal model. Leong and Mihalcea (2011) have developed an approach that has been later referred to as scoring-level fusion. They used distinct text-based and image-based models to compute word relatedness, and combined similarity scores from these monomodal models by taking the sum or harmonic mean. The resulting hybrid similarity measures outperformed both text- and image-based models used in isolation. A more extensive approach for combining multimodal data into a single model is feature-level fusion, in which semantic representations from different modalities are combined in the feature space to create multimodal vectors (Bruni, Tran, and Baroni 2014; Kiela and Bottou 2014). Features from different modalities can be concatenated into a single matrix and projected onto a common space using a form of dimensionality reduction. Even more flexible construction of multimodal features can be achieved using deep learning methods. By concatenating the features from multiple modalities and feeding them to a supervised classifier, semantic representations can be learned that are fit to a specific task. This type of fusion has been implemented using various learning structures such as traditional neural networks (Poria, Cambria, and Gelbukh 2015) and Deep Belief networks consisting of stacked Restricted Boltzmann Machines (Kim, Lee, and Provost 2013; Ngiam et al. 2011).

# 3 Methods

## 3.1 String metric learning for kernel regression

Following the methodology of Gutierrez et al (2016), we use a kernel regression framework to analyze form-meaning systematicity. Kernel regression is a nonparametric supervised learning technique that is widely used for pattern detection problems. Data samples are defined by predictor variables as well as target variables. Target variable values for individual data samples are predicted based on their distance in predictor variables to other data samples, for which the target variables are known. This enables the model to capture local structures in the data, in contrast to parametric models that generally provide a more global fit (Takeda, Farsiu, and Milanfar 2007).

We implement the linear Nadaraya-Watson estimator (Nadaraya 1964). Given a set of N data points $\{x_i\}_{i=1}^N$ with target values $\{y_i\}_{i=1}^N$, the Nadaraya-Watson estimator for a data sample $x_j$ defined as follows:

$$\hat{y}(x_j) = \frac{\sum_{i \neq j} k_{ij} * y_j}{\sum_{i \neq j} k_{ij}}, \tag{1}$$

where $k_{ij}$ is the kernel between data points i and j, computed using a kernel function that penalizes distance in predictor variables between two samples. We implement the following exponential kernel function:

$$k(x_i, x_j) = exp(-d(x_i, x_j)/h). \tag{2}$$

The variable $h$ specifies a bandwidth that determines the radius of the neighborhood in which data samples effectively contribute to each other's prediction. The distance metric $d$ defines the distance in predictor variables between data samples.

We use this framework to predict semantic vectors of words based on their strings, which function as predictor variables. Following previous research (Nosofsky 1986; Gutiérrez, Levy, and Bergen 2016), we use Levenshtein edit-distance as a distance measure (Levenshtein 1966). The distance between two strings is measured as the minimum number of edits needed to transform one string into another. An edit is a mutation, insertion or deletion of one letter. We choose to use Levenshtein edit-distance over phonetic distance to avoid noise resulting from sound shifts between English dialects. In order to avoid this noise, a phonetic apprach would need to involve a corpus curated to include texts from a single dialect. Since previous research has indicated that string-edits differ in their semantic relevance (Gil et al. 2005; Gutiérrez, Levy, and Bergen 2016), we optimize a weight matrix for string-edits by minimizing the mean squared error (MSE) of kernel regression using gradient descent. The error is computed as:

$$\mathcal{L} = \sum_{i=1}^{N} ((y_i - \hat{y}_i)^T (y_i - \hat{y}_i)). \tag{3}$$

Following Gutierrez et al. (2016), we refer to this method as *string metric learning for kernel regression* (SMLKR). Since all weights are set to minimize the MSE, SMLKR computes weights that in the process optimize the bandwidth variable *h*.

To implement SMLKR, the edit paths between strings are stored as vectors $V$, in which each dimension represents one type of string edit (e.g. substituting an 'a' for a 'b', or deleting a 't'), as shown in Figure 1. Substitutions are represented symmetrically so that opposite substitutions share the same semantic significance. Furthermore, the weights are bounded between 0 and infinity, since negative weights for string-edits imply that an edit negatively contributes to the distance between two strings (Gutiérrez, Levy, and Bergen 2016).

The edit-vectors *V* are multiplied with a weight-vector *W* of the same size *S*, resulting in a single weighted edit-distance for each pair of strings. This can be formally stated as follows:

$$d(s_i, s_j) = \sum_{s=1}^{S} (W_s * V_{ijs}) = W^T V_{ij}. \tag{4}$$

We can now compute the gradient of MSE as follows:

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial \hat{y}_i} * \frac{\partial \hat{y}_i}{\partial W}. \tag{5}$$

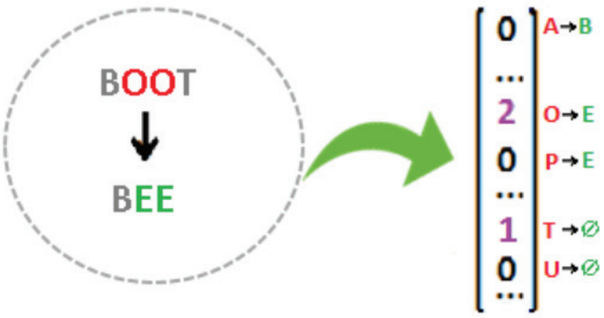

Figure 1: Mutation vector between 'boot' and 'bee' (Gutiérrez, Levy, and Bergen 2016)

Where the partial derivatives are:

$$\frac{\partial \mathcal{L}}{\partial y_i} = \frac{2}{N} * \sum_{i=1}^{N} (y_i - \hat{y}_i) \tag{6}$$

$$\frac{\partial \hat{y}_i}{\partial W} = \frac{\sum_{j \neq i} (y_j - \hat{y}_i)^T k_{ij} v_{ij}}{\sum_{j \neq i} k_{ij}}. \tag{7}$$

**Analyzing form-meaning systematicity** After we optimize the weights for string-edits we correlate the weighted Levenshtein distances with semantic distances to quantify the significance of form-meaning systematicity in the lexicon. We compute correlations using the Mantel test (Mantel 1967). This test computes the Pearson correlation between two distance-matrices in which entries on the same indexes are paired. Subsequently, both matrices are subjected to random permutations, after which the same correlation is computed. The proportion of permuted matrix-pairs that display a higher correlation than the initial two matrices is computed as p-value. This represents the probability that the measured correlation is found in a language corpus under the null hypothesis that form-meaning relationship is arbitrary.

The optimized distance-metric is also used to predict semantic representations of words based on their strings. Following Gutierrez et al. (2016), we take the extent to which the semantic vector of a word can be predicted as a measure of its systematicity. Under this assumption, we analyze the kernel regression error for words containing possible phonaesthemes. We compare the average error of all words belonging to a single phonaestheme, with the average error of 1000 random samples of equal size. The portion of random sets with a lower mean error is assigned to the investigated phonaestheme as a p-value. This represents the probability of a set of words displaying the identified systematicity, under the null hypothesis that phonaesthemes do not exist in the lexicon (Gutiérrez, Levy, and Bergen 2016).

## 3.2 Semantic representations

**Lexicon** We construct semantic representations for all words in a lexicon of English monomorphemes. We use the same lexicon as Gutierrez et al. (2016), which has been constructed by cross-referencing monomorphemic English words in the CELEX lexical database (Baayen, Piepenbrock, and Gulikers 1996) with monomorphemic words in the Oxford English Dictionary Online (Simpson, Weiner, and others 1989). Remaining polymorphemic words, place names, demonyms, spelling variants and proper nouns have been removed. Words that were not among the 40,000 most frequent non-filler word types were excluded. We removed words for which no text- or image-based semantic representation was available, as they were absent in the datasets the models were trained on. This resulted in a final list of 4479 monomorphemic words (out of the 4958 used by Gutierrez et. al).

**Text-based semantic vectors** We used skip-gram with negative sampling (Mikolov et al. 2013) trained on the Google News dataset as our text-based model. The model is freely available as part of the Word2Vec system release[1], along with 300-dimensional vector representations for 3 million words and phrases.

**Visual semantic vectors** We constructed our image-based semantic vectors using the MMFeat toolkit (Kiela 2016), which relies on the Caffe deep learning framework (Jia et al. 2014). We first retrieved 10 images for each word in our lexicon using Bing Image Search. We then extracted an embedding for each of the images from a deep convolutional neural network that was trained on the ImageNet classification task (Russakovsky et al. 2015). The network architecture consisted of five convolutional layers, followed by two fully connected rectified linear unit (ReLU) layers and a softmax layer for classification. It was trained with a multinomial logistic regression objective:

$$J(\theta) = -\sum_{i=1}^{D} \sum_{k=1}^{K} \mathbf{1}\{y^{(i)} = k\} \log \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})} \tag{8}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $D$ – the number of examples and $K$ – the number of classes. To obtain an embedding for a given image we performed a forward pass through the network and extracted the 4096-dimensional fully connected layer that precedes the softmax layer as the representation of that image. We aggregated the visual representations of the 10 images retrieved for each word by taking their average. Such a transfer learning approach has the advantage of using a large dataset of manually annotated images to train the model, while ensuring that sufficient visual data is available for the words in the lexicon.

**Multimodal models** We investigated three methods for combining text- and image-based semantic vectors into a multimodal model.

*Scoring-level fusion:* Our first approach is based on scoring-level fusion (Leong and Mihalcea 2011). We perform kernel regression separately on the linguistic and visual models, and compute the semantic distance between two words

[1] http://code.google.com/archive/p/word2vec/

as a weighted average of the cosine distances between the two respective vectors in the linguistic and the visual models. We correlate this multimodal semantic distance with unweighted Levenshtein distance as well as weighted Levenshtein distance. The weighted average of semantic distances and the weighted average of optimized Levenshtein distances are computed using the same parameter $\alpha$:

$$\alpha * \text{linguistic distance} + (1 - \alpha) * \text{visual distance} \quad (9)$$

The resulting correlation is computed using different values for $\alpha$, after which the weighting factor that results in the highest correlation is identified as optimal. This optimal weighting factor for our two monomodal models was 0.75. This shows that the two models are complementary. Furthermore, $\alpha$ stays the same under weight-optimization, confirming that it is predominantly dependent on the extent to which the information conveyed in both monomodal models is complementary.

*Feature-level fusion:* Secondly, we implement feature-level fusion by first concatenating the linguistic and visual semantic representations and then running kernel regression on this multimodal model. This approach has the benefit over the scoring-level fusion of explicitly training weights to optimize the prediction of multimodal semantic representations. However, a problem arises with simple concatenation since the dimensionality of the image-based vectors (4096) is of a different order of magnitude than the dimensionality of the text-based vectors (300). We therefore normalize the linguistic and visual models separately before concatenation.

*Neural network fusion:* Thirdly, we trained a neural network with a Siamese architecture to extract multimodal features, similar to the approach of Poria et al. (2015). We use the concatenation of the linguistic and visual feature vectors as input, and randomly pair words in the lexicon. We then use the Levenshtein edit distances between the two words in the pair as values to be predicted. The concatenated representation for each word in the pair is fed to a separate branch of the Siamese network. The first three layers of both branches are fully connected ReLU layers. Each layer shares the same weights over both branches. After the third layer, we concatenate the vectors in both branches and predict the edit distance between the two words.

We train the network on the entire lexicon using MSE as the loss function. Subsequently, we perform a forward pass through the trained network for all words in the lexicon and extract the 300 dimensional final layer — i.e. the last layer in the word's branch before concatenation — as a multimodal representation of that word. The neural network architecture is shown in Figure 2.

## 4 Experiments and Results

We use SMLKR to optimize edit weights for all monomodal and multimodal models. We initialize edit weights to 1 and optimize them by minimizing the MSE until convergence. We perform two experiments on all models. Firstly, we correlate edit-distance with semantic distance to measure form-meaning systematicity over the lexicon as a whole. Secondly, we analyze the predictability of the semantic vectors of individual words from their strings, to identify phonaesthemes.

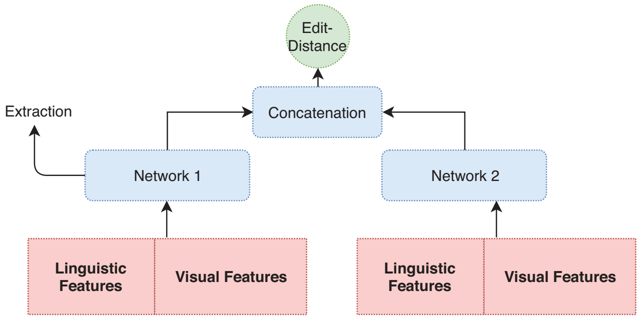

Figure 2: The neural network fusion process

| Model | Correl. | p-value |
|---|---|---|
| Text-based | 0.0383 | 0.001 |
| Image-based | 0.0243 | 0.007 |
| Scoring-level fusion | **0.0420** | 0.001 |
| Multimodal concat | 0.0376 | 0.001 |
| Neural Network fusion | 0.0266 | 0.001 |

Table 1: Correlations between weighted edit distances and cosine distances for all models

| Model | Correl. | p-value |
|---|---|---|
| Text-based | 0.0362 | 0.001 |
| Image-based | 0.0198 | 0.025 |
| Scoring-level fusion | **0.0401** | 0.001 |
| Multimodal concat | 0.0351 | 0.001 |
| Neural Network fusion | 0.0175 | 0.001 |

Table 2: Correlations between unweighted edit distances and cosine distances for all models

### 4.1 Lexicon-wide form-meaning systematicity

Table 1 shows correlations between weighted edit distances and semantic cosine distances for all monomodal and multimodal models. All correlations have been computed using the Mantel permutation test. All models exhibit a statistically significant correlation between form and meaning, which shows that form-meaning systematicity exists in the lexicon. All correlations are relatively small (<0.05). This is to be expected under the assumption that concentrated systematic clusters exist as exceptions to the overall arbitrariness in language (Gutiérrez, Levy, and Bergen 2016). Optimizing the edit weights has improved the correlation for all models, as compared to using unweighted distances, as shown in Table 2. This shows that string edits indeed differ in their semantic relevance. Interestingly, we observe some systematicity when using the visual model alone, which lends support to the existence of iconicity as a relationship between word form and the visual properties of the referent.

The multimodal model constructed using scoring-level fusion significantly outperforms the text-based model used in isolation. This further shows that incorporating visual fea-

| Ph | P-value | Systematic Words |
|---|---|---|
| sn- | <0.0001 | sneeze, sniff, snore, snort, snout |
| mu- | <0.0001 | muck, mushroom, mush, musk |
| tw- | <0.0001 | twitch, twang, twinkle, twit |
| sq- | 0.0032 | squabble, squeak, squeal, squirt |
| pe- | 0.0079 | pea, peach, pear, pearl, pebble |
| bu- | 0.0087 | buff, buffalo, bull, bully |
| sw- | 0.0173 | swirl, swish, swipe, swerve |
| cr- | 0.0137 | crab, crawl, creep, crouch |

Table 3: Phonaesthemes (text-based model)

tures increases the level of form-meaning systematicity that can be found in language.

In order to test the significance of the pairwise differences between correlations produced by different models, we used a bootstrap test (Pernet, Wilcox, and Rousselet 2013). For each pair of models, we begin by sampling words from the lexicon with replacement, and recompute the cosine distances and edit distances for this resampled lexicon. We then compute the correlations between these resampled distances, and test how often we observe the same relationship between the correlations for the different models in Tables 1 and 2. This procedure is repeated for 2000 iterations, yielding bootstrapped p-values for the significance of the differences between each pair of models. The differences between all model pairs were significant with $p < 0.01$, except for that between the weighted multimodal concatenation and text-based models ($p = 0.14$), and between the weighted and unweighted visual models ($p = 0.06$).

The multimodal concatenation and neural network fusion models do not display more systematicity than the text-based model, suggesting that scoring-level fusion is a superior multimodal approach in this task. However, since scoring-level fusion does not involve constructing multimodal vectors it is unsuitable for qualitative analyses, which we conduct below.

## 4.2 Phonaesthemes

We analyze to what extent form-meaning systematicity is localized in phonaesthemes by comparing the average kernel regression error for each phonosemantic cluster with 1000 random samples in the lexicon as described above. For this analysis, we use weighted edit distances which have been optimized for each individual model. Tables 3, 4, 5 and 6 show the two letter onsets (i.e. phonaesthemes) that display the most form-meaning systematicity for each investigated semantic model with the exception of scoring-level fusion. The p-value represents the probability that the identified systematicity for the relevant cluster occurs under the null-hypothesis that systematicity is not localized in phonosemantic clusters. The third column lists the most systematic words whithin these clusters.

The analysis of the most systematic words for each phonaestheme shows that words in many phonaesthemes resemble each other in their meaning. For instance, in Table 3 ’sn-’ refers to the nose, ’mu-’ refers to the ground and dirt, ’tw-’ refers to small motion and ’sw-’ refers to larger motion. We found a total of 22 phonaesthemes with $p < 0.05$ using the text-based model.

| Ph | P-value | Systematic Words |
|---|---|---|
| jo- | 0.0113 | join, joint, joke, jolly, jolt, joy |
| hu- | 0.0125 | hubbub, huddle, hustle, hubris |
| si- | 0.0202 | sick, sigh, sin, silly, simple |
| id- | 0.0209 | idea, idiom, idiot, idle, idol, idyll |
| pr- | 0.0230 | pray, preach, pride, priest |
| fa- | 0.0270 | fail, faint, fallacy, farce, falter |

Table 4: Phonaesthemes (image-based model)

| Ph | P-value | Systematic Words |
|---|---|---|
| sn- | 0.0020 | sneeze, sniff, snore, snort, snout |
| hu- | 0.0037 | hubbub, hustle, hubris |
| id- | 0.0089 | idea, ideion, idiot, idol, idyll |
| cr- | 0.0116 | crab, crawl, creep, crouch |
| fl- | 0.0165 | flash, fleck, flee, flick, flinch |
| si- | 0.0286 | sight, sign, silhouette, simulate |
| fa- | 0.0489 | fail, faint, fallacy, fake, famine |
| jo- | 0.0492 | job, joke, jolly, joy, jolt, jot |

Table 5: Phonaesthemes (multimodal concatenation)

Systematicity is far less clustered in the image-based and multimodal models. The image-based model identified 15 phonaesthemes with $p < 0.05$. The multimodal concatenation and neural network fusion models respectively show 10 and 7 phonaesthemes with $p < 0.05$. Furthermore, the image-based and multimodal models identify different phonaesthemes than the text-based model. The text-based models tend to capture more concrete meaning similarities between words. The image-based and multimodal models, on the other hand, capture similarities of a more abstract nature. The image-based model for instance, finds ’jo-’, which refers to joy, ’hu-’, which refers to disorganization, ’pr-’, which refers to high esteem and religion and ’fa-’, which has a negative connotation. The multimodal concatenation and neural network fusion models discover phonaesthemes found by the text-based model as well as phonaesthemes found by the image-based model. However, the results from both multimodal models predominantly resemble the results found using the image-based model. In addition, both multimodal models identify phonaesthemes that are not found using either of the two monomodal models. The multimodal concatenation model identifies ’fl-’ which refers to quick and small motion and ’si-’ which refers to vision. The neural network fusion model identifies ’bl-’ which refers to whiteness and cleanliness, and ’le-’ which refers to education.

We apply an etymological analysis to the novel phonosemantic clusters identified by our image-based and multimodal models to determine the relative origins of words within these clusters. Interestingly, many phonaesthemes originate from phonosemantic clusters in a parent language. For instance, ’pray’ ’preach’, and ’priest’ originate from Latin roots with a meaning similar to their English descendents: ’precari, praediceare’ and ’presbyter’ respectively. Similarly, ’crab’, ’crawl’ and ’creep’ stem from the old

| Ph | P-value | Systematic Words |
|---|---|---|
| hu- | 0.0173 | huge, hulk, humiliate, humble ... |
| bl- | 0.0289 | blanch, blank, bless, blight, bliss |
| le- | 0.0300 | learn, lethargy, lesson, letter |
| id- | 0.0352 | idea, ideion, idiot, idle, idol, idyll |

Table 6: Phonaesthemes (neural network fusion)

Norse 'krabbi', 'krafla' and 'kjurpa', which again have a similar meaning. However, some phonosemantic clusters contain words with varying origins of which the meanings converged more recently. For instance, 'sign' and 'simulate' stem from the Latin 'signum' and 'simulare' which mean 'sign/mark' and 'imitate' respectively, while 'sight' stems from the Proto-Germanic 'sekh' directly referring to vision. Furthermore, 'bless' originates from the Proto-Germanic 'blodison' meaning 'to hallow with blood', while 'bliss' stems from the Proto-Germanic 'blithsjo' meaning 'kind' (Harper 2001).

These different cases are to be expected under the assumption that form-meaning systematicity is a universal phenomenon across languages which are itself subject to constant change. Some correlations between form and meaning are the direct result of systematicity in earlier linguistic systems, while other semantic clusters are formed from independent origins such as 'blodison' and 'blithsjo' that converge both in form and meaning during the process of linguistic development.

## 5 Discussion

Our findings support the existence of form-meaning systematicity in the English language. All monomodal and multimodal models exhibit statistically significant correlations between form and meaning. However, all correlations are relatively small (<0.05). This shows that while form-meaning systematicity does exist, the larger part of form-meaning mapping is arbitrary. We have demonstrated that the multimodal model displays a higher correlation (0.0420) between form and meaning than both text-based and image-based models in isolation (0.0383 and 0.0243 respectively). This difference is statistically significant with $p < 0.01$. This shows that the text- and image-based models convey complementary semantic information, and that it is beneficial to combine these modalities when investigating form-meaning systematicity in language. These results further support one of the hypotheses behind iconicity, i.e. that word forms may be motivated by the visual aspects of the referent.

The concatenation and neural network fusion models do not reach the same performance as the scoring-level fusion model in this task, indicating that these models do not fully utilize the potential of the two modalities. One possible explanation for this is the large difference in dimensionality between the text- and image-based semantic vectors. Whilst these dimensionalities have been shown to be optimal in the respective NLP and computer vision tasks, it may be the case that the higher dimensionality of the image-based vectors biases the multimodal model towards visual information, disregarding important aspects of the meaning derived from textual data. Furthermore, we have trained the neural network fusion model by predicting the edit distances between words based on their semantic representations. The motivation for this was to train multimodal semantic representations that are specialised for the task of identifying form-meaning systematicity. It is possible, however, that the task of predicting edit distances is insufficient or unsuitable for optimizing semantic representations. After all, relations between form and meaning are not evident throughout the entire lexicon. A semantic classification task, such as word categorization or lexical entailment, might prove to be more effective in training multimodal representations, which we intend to investigate in our future work.

Nonetheless, the qualitative analysis of phonaesthemes has shown the value of all multimodal models. All models have identified phonaesthemes, further supporting the hypothesis that systematicity is in fact concentrated in localized clusters. As noted earlier, the multimodal models are predominantly influenced by the image-based model. Yet, each multimodal model has identified systematic clusters that were not found by any of the monomodal models or another multimodal model. This suggests that each method for combining information from different modalities captures a slightly different semantic representation, and that each is able to identify different systematic structures.

## 6 Conclusion

We have presented the first multimodal approach to linguistic arbitrariness. Using data-driven linguistic and visual models, we have provided further evidence for the claim that systematic relations between form and meaning exist in the English language. Our results demonstrate that incorporating visual features can substantially increase the level of form-meaning systematicity found in language. Additionally, our qualitative analysis of phonaesthemes has shown that multimodal models can be used to identify systematic relations that are not found using any monomodal models.

Using multimodal distributional semantics to analyze form-meaning systematicity opens up new research avenues, which can lead to a better understanding of how the vocabularies of human languages have evolved as a way to convey meaning. Much progress can still be made in the techniques used to integrate information from multiple modalities to study form-meaning systematicity. Another interesting future research direction would be to experiment with audio as well as visual data, capturing both iconicity and onomatopoeic relations between form and meaning. Finally, it is essential to apply the methods to languages other than English, investigating to what extent form-meaning systematicity is a universal phenomenon across languages.

## References

Baayen, R. H.; Piepenbrock, R.; and Gulikers, L. 1996. Celex2. *Philadelphia: Linguistic Data Consortium*.

Barsalou, L. W. 2008. Grounded Cognition. *Annu Rev Psychol* 59:617–645.

Bruni, E.; Tran, N. K.; and Baroni, M. 2014. Multimodal distributional semantics. *JAIR* 49:1–47.

De Saussure, F., and Baskin, W. 1916. Course in general linguistics. *London: Duckworth*.

Fincher-Kiefer, R., and D'Agostino, P. R. 2004. The role of visuospatial resources in generating predictive and bridging inferences. *Discourse Processes* 37(3):205–224.

Gasser, M. 2004. The origins of arbitrariness in language. *Proceedings of CogSci, 26(26)* 30(5):434–439.

Gernsbacher, M. A.; Varner, K. R.; and Faust, M. E. 1990. Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(3):430.

Gil, Y.; Motta, E.; Benjamins, V. R.; and Musen, M. A. 2005. The semantic web-iswc 2005. In *4th ISWC*, 6–10.

Gutiérrez, E.; Levy, R.; and Bergen, B. K. 2016. Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression. *Acl* (1984):2379–2388.

Harper, D. 2001. Online etymology dictionary.

Harris, Z. 1970. Distributional structure. In *Papers in structural and transformational linguistics*. Springer. 775–794.

Hockett, C. F., and Hockett, C. D. 1960. The origin of speech. *Scientific American* 203(3):88–97.

Hutchins, S. S. 1999. The psychological reality, variability, and compositionality of english phonesthemes.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings ACM(22)*, 675–678.

Kiela, D., and Bottou, L. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. *In Proceedings of EMNLP 2014* 36–45.

Kiela, D.; Rimell, L.; Vulić, I.; and Clark, S. 2015. Exploiting image generality for lexical entailment detection. In *Proceedings of ACL 2014*.

Kiela, D. 2016. MMF EAT : A Toolkit for Extracting Multi-Modal Features. *Proceedings of ACL 2016* 55–60.

Kim, Y.; Lee, H.; and Provost, E. M. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of ICASSP 2013*, 3687–3691. IEEE.

Leong, C., and Mihalcea, R. 2011. Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness. *Ijcnlp* 1403–1407.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, 707–710.

Liu, N. F.; Levow, G.-A.; and Smith, N. A. 2018. Discovering phonesthemes with sparse regularization. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, 49–54.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research* 27(Part 1):209–220.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Monaghan, P.; Shillcock, R. C.; Christiansen, M. H.; and Kirby, S. 2014. How arbitrary is language? *PTRS B* 369.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9(1):141–142.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of ICML 11)*, 689–696.

Nosofsky, R. M. 1986. Attention, similarity, and the identification–categorization relationship. *J Exp Psychol* 115(1):39.

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

Otis, K., and Sagi, E. 2008. Phonaesthemes: A corpus-based analysis. In *Proceedings of the 11th CogSci*, volume 30.

Pernet, C. R.; Wilcox, R. R.; and Rousselet, G. A. 2013. Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. *Frontiers in psychology* 3:606.

Poria, S.; Cambria, E.; and Gelbukh, A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of EMNLP 2015*, 2539–2544.

Roller, S., and Schulte im Walde, S. 2013. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of EMNLP 2013*, 1146–1157.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115(3):211–252.

Shillcock, R.; Kirby, S.; McDonald, S.; and Brew, C. 2001. Filled pauses and their status in the mental lexicon. In *ISCA ITRW on Disfluency in Spontaneous Speech*.

Shutova, E.; Kiela, D.; and Maillard, J. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *NAACL HLT 2016*, 160–170.

Simpson, J.; Weiner, E. S.; et al. 1989. Oxford english dictionary online. *Oxford: Clarendon Press* 6:2008.

Takeda, H.; Farsiu, S.; and Milanfar, P. 2007. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing* 16(2):349–366.

Wittgenstein, L. 1953. 'philosophical investigations: Translated by gem˜ anscombe. oxford. blackwell.

Zwaan, R. A., and Madden, C. J. 2005. Embodied sentence comprehension. *Grounding cognition: The role of perception and action in memory, language, and thinking* 224–245.