Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression

E. Darío Gutiérrez
UC San Diego
edg@icsi.berkeley.edu

Roger Levy
MIT
rplevy@mit.edu

Benjamin K. Bergen
UC San Diego
bkbergen@ucsd.edu

Abstract

Arbitrariness of the sign—the notion that the forms of words are unrelated to their meanings—is an underlying assumption of many linguistic theories. Two lines of research have recently challenged this assumption, but they produce differing characterizations of non-arbitrariness in language. Behavioral and corpus studies have confirmed the validity of localized form-meaning patterns manifested in limited subsets of the lexicon. Meanwhile, global (lexicon-wide) statistical analyses instead find diffuse form-meaning systematicity across the lexicon as a whole.

We bridge the gap with an approach that can detect both local and global form-meaning systematicity in language. In the kernel regression formulation we introduce, form-meaning relationships can be used to predict words' distributional semantic vectors from their forms. Furthermore, we introduce a novel metric learning algorithm that can learn weighted edit distances that minimize kernel regression error. Our results suggest that the English lexicon exhibits far more global form-meaning systematicity than previously discovered, and that much of this systematicity is focused in localized form-meaning patterns.

1 Introduction

Arbitrariness of the sign refers to the notion that the phonetic/orthographic forms of words have no relationship to their meanings (de Saussure, 1916). It is a foundational assumption of many theories of language comprehension, production, acquisition, and evolution. For instance, Hockett's (1960) influential enumeration of the design features of human language ascribes a central role to arbitrariness in enabling the combination and recombination of phonemic units to create new words. Gasser (2004) uses simulations to show that for large vocabularies, arbitrary form-meaning mappings may provide an advantage in acquisition. Meanwhile, modular theories of language comprehension rely upon the duality of patterning to support the independence of the phonetic and semantic aspects of language comprehension (Levelt et al., 1999). Quantifying the extent to which the arbitrariness principle actually holds is important for understanding how language works.

Language researchers have long noted exceptions to arbitrariness. Most of these are patterns that occur in some relatively localized subset of the lexicon. These patterns are sub-morphemic because, unlike conventional morphemes, they cannot combine reliably to produce new words. *Phonaesthemes* (1930) are one example. A phonaestheme is a phonetic cluster that recurs in many words that have related meanings. One notable phonaestheme is the onset *gl-*, which occurs at the beginning of at least 38 English words relating to vision: *glow*, *glint*, *glaze*, *gleam*, etc. (Bergen, 2004). At least 46 candidate phonaesthemes have been posited in the linguistics literature, according to a list compiled by Hutchins (1998). *Iconicity* is another violation of arbitrariness that can lead to non-arbitrary local regularities. Iconicity occurs when the form of a word is transparently motivated by some perceptual aspect of its referent. Consequently, when several referents share perceptual features, their associated word-forms would tend to be similar as well (to the extent that they are iconic). For instance, Ohala (1984) conjectures that vowels with high acoustic frequency tend to associate with smaller items while vowels with low acoustic frequency

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2379–2388,
Berlin, Germany, August 7-12, 2016.

tend to associate with larger items, due to the experiential link between vocalizer size and frequency. Systematic iconicity is also manifested in sets of onomatopoeic words that echo similar sounds (e.g., *clink*, *clank*). Although these exceptions to non-arbitrariness differ, in each case, specific form-meaning relationships emerge in a subset of the lexicon. We will refer to all such specific localized form-meaning patterns as *phonosemantic sets*.

In recent decades, behavioral and corpus studies have empirically confirmed the psychological reality and statistical reliability of many phonosemantic sets that had previously been identified by intuition and observation. Various candidate phonaesthemes have significant effects on reaction times during language processing tasks (Hutchins, 1998; Magnus, 1998; Bergen, 2004). Sagi and Otis (2008) test the statistical significance of the 46 candidates in Hutchins's (1998) list, and find that 27 of them exhibit more within-category distributional semantic coherence than expected by chance. These results have been replicated using other corpora and distributional semantic models (Abramova et al., 2013). Klink (2000) shows that sound-symbolic attributes such as those proposed by Ohala (1984) are associated with human judgments about nonwords' semantic attributes, such as smallness or beauty. Using a statistical corpus analysis and WordNet semantic features, Monaghan et al. (2014a) examine a similar hypothesis space of sound-symbolic phonological and semantic attributes, and reach similar conclusions.

While these localized studies support the existence of some islands of non-arbitrariness in language, their results do not address how pervasive non-arbitrariness is at the global level—that is, in the lexicon of a language as a whole. After all, some seemingly non-arbitrary local patterns can be expected to emerge merely by chance. How can we measure whether local phonosemantic patterning translates into global *phonosemantic systematicity*–that is, strong, non-negligible lexicon-wide non-arbitrariness? Shillcock et al. (2001) introduce the idea of measuring phonosemantic systematicity by analyzing the correlation between phonological edit distances and distributional semantic distances. In a lexicon of monomorphemic and monosyllabic English words, they find a small but statistically significant correlation between these two distance measures. Monaghan et al. (2014b) elaborate on this methodology, showing that the statistical effect is robust to different choices of form-distance and semantic-distance metrics. They also look at the effect of leaving out each word in the lexicon on the overall correlation measure; from this, they derive a phonosemantic systematicity measure for each word. Interestingly, they find that systematicity is diffusely distributed across the words in English in a pattern indistinguishable from random chance. Hence, they conclude that "systematicity in the vocabulary is not a consequence of small clusters of sound symbolism." This line of work provides a proof-of-concept that it is possible to detect the phonosemantic systematicity of a language, and confirms that English exhibits significant phonosemantic systematicity.

Broadly speaking, both the localized tests of individual phonosemantic sets and the global analyses of phonosemantic systematicity challenge the arbitrariness of the sign. However, they attribute responsibility for non-arbitrariness differently. The local methods reveal dozens of specific phonosemantic sets that have strong, measurable behavioral effects and statistical signatures in corpora. Meanwhile, the global methods find small and diffuse systematicity. How can we reconcile this discrepancy?

Original Contributions. We attempt to bridge the gap with a new approach that builds off of previous lexicon-wide analyses, making two innovations. The first addresses the concern that the lexicon-wide methods currently in use may not be well suited to finding local regularities such as phonosemantic sets, because they make the assumption that systematicity exists only in the form of a global correlation between distances in form-space and distances in meaning-space. Instead, we model the problem using kernel regression, a non-parametric regression model. Crucially, in kernel regression the prediction for a point is based on the predictions of neighboring points; this enables us to conduct a global analysis while still capturing local, neighborhood effects. As in previous work, we represent word-forms by their orthographic strings, and word-meanings by their semantic vector representations as produced by a distributional semantic vector space model. The goal of the regression is then to learn a mapping from string-valued predictor variables to vector-valued target variables that minimizes regression

error in the vector space. Conveniently, our model allows us to produce predictions of the semantic vectors associated with both words and nonwords.

Previous work may also underestimate systematicity in that it weights all edits (substitutions, insertions, and deletions) equally in determining edit distance. A priori, there is no reason to believe this is the case—indeed, the work on individual phonosemantic sets suggests that some orthographic/phonetic attributes are more important than others for non-arbitrariness. To address this, we introduce String-Metric Learning for Kernel Regression (SMLKR), a metric-learning algorithm that is able to learn a weighted edit distance metric that minimizes the prediction error in kernel regression.

We find that SMLKR enables us to recover more systematicity from a lexicon of monomorphemic English words than reported in previous global analyses. Using SMLKR, we propose a new measure of per-word phonosemantic systematicity. Our analyses using this systematicity measure indicate that specific phonosemantic sets do contribute significantly to the global phonosemantic systematicity of English, in keeping with previous local-level analyses. Finally, we evaluate our systematicity measure against human judgments, and find that it accords with raters' intuitions about what makes a word's form well suited to its meaning.

2 Background & Related Work

2.1 Previous Approaches to Finding Lexicon-Wide Systematicity

Measuring Form, Meaning, and Systematicity. To our knowledge, all previous lexicon-level analyses of phonosemantic systematicity have used variations of the method of Shillcock et al. (2001). The inputs for this method are form-meaning tuples $(\mathbf{y}_i, s_i)$ for each word i in the lexicon, where $\mathbf{y}_i$ is the vector representation of the word in a distributional semantic model, and s_i is the string representation of the word (phonological, phonemic, or orthographic). Semantic distances are measured as cosine distances between the vectors of each pair of words. Shillcock et al. (Shillcock et al., 2001) and Monaghan et al. (Monaghan et al., 2014b) measure form-distances in terms of edit distance between each pair of strings. In addition Monaghan et al. (2014b) and Tamariz (2006) study distance measures based on a selected set binary phonological features, with similar results. Phonosemantic systematicity is then measured as the correlation between all the pairwise semantic distances and all the pairwise string distances.

Hypothesis Testing. In this line line of work, statistical significance of the results is assessed using the Mantel test, a permutation test of the correlation between two sets of pairwise distances (Mantel, 1967). The test involves randomly shuffling the assignments of semantic vectors to wordstrings in the lexicon. We can think of each form-meaning shuffle as a member of the set of all possible lexicons. Next, the correlation between the semantic distances and the string distances is computed under each reassignment. An empirical p-value for the true lexicon is then derived by performing many shufflings, and comparing the correlation coefficients measured under the shuffles to the correlation coefficient measured in the true lexicon. Under the null hypothesis that form-meaning assignments are arbitrary, the probability of observing a form-meaning correlation of at least the magnitude actually observed in the true lexicon is asymptotically equal to the proportion of reassignments that produce greater correlations than the true lexicon.

Previous Findings. Shillcock et al. (2001) find a statistically significant correlation between semantic and phonological edit distances in a lexicon of the 1733 most frequent monosyllabic monomorphemic words in the BNC. Tamariz (2008) extends these results to Spanish data, looking only at words with one of three consonant-vowel (CV) structures (CVCV, CVCCV, and CVCVCV). (2001), Monaghan et al. (2014b) derive a list of 5138 monomorphemic monosyllabic words and a list of 5604 monomorphemic polysyllabic from the CELEX database (Baayen et al., 1996), and find significant form-meaning correlations in both.

2.2 Kernel Regression

In contrast to previous studies, we study form-meaning systematicity using a kernel regression framework. Kernel regression is a nonparametric supervised learning technique that is able to learn highly nonlinear relationships between predictor variables and target variables. Rather than assuming any particular parametric relationship between the predictor and target variables, kernel regression assumes only that the value of the target vari-

able is a smooth function of the value of the predictors. In other words, given a new point in predictor space, the value of the target at that point can reasonably be estimated by the value of the targets at points that are nearby in the predictor space. In this way, kernel regression is analogous to an exemplar model. We performed kernel regression on our lexicon using the Nadaraya-Watson estimator (Nadaraya, 1964). Given a data set $\mathcal{D}$ of vector-valued predictor variables $\{\mathbf{x}_i\}_{i=1}^N$, and targets $\{\mathbf{y}_i\}_{i=1}^N$, the Nadaraya-Watson estimator of the target for sample i is

$$\hat{\mathbf{y}}_i = \hat{\mathbf{y}}(\mathbf{x}_i) = \frac{\sum_{j \neq i} k_{ij} \mathbf{y}_j}{\sum_{j \neq i} k_{ij}}, \quad (1)$$

where k_{ij} is the *kernel* between point i and point j. A commonly used kernel is the exponential kernel:

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-d(\mathbf{x}_i, \mathbf{x}_j)/h),$$

where $d(\cdot, \cdot)$ is a distance metric and h is a *bandwidth* that determines the radius of the effective neighborhood around each point that contributes to its estimate. For our purposes we use the Levenshtein string edit distance metric (Levenshtein, 1966). The Levenshtein edit distance between two strings is the minimum number of edits needed to transform one string into the other, where an edit is defined as the insertion, deletion, or substitution of a single character. Using this edit distance and semantic vectors derived from a distributional semantic model, the Nadaraya-Watson estimator can estimate the position in the semantic vector space for each word in the lexicon. The exponential edit distance kernel has been useful for modeling behavior in many tasks involving word similarity and neighborhood effects; see, for example the Generalized Context Model (Nosofsky, 1986), which has been applied to word identification, recognition, and categorization, to inflectional morphology, and to artificial grammar learning (Bailey and Hahn, 2001).

2.3 Metric Learning for Kernel Regression

In kernel regression, the bandwidth h of the kernel function must be fine-tuned by testing out many different bandwidths. Moreover, for many tasks there is no reason to assume that all of the dimensions of a vector-valued predictor are equally important. This is problematic for conventional kernel regression, as the quality of its predictions is wholly reliant on the appropriateness of the given distance metric.

Weinberger and Tesauro (2007) introduce metric learning for kernel regression (MLKR), an algorithm that can learn a task-specific Mahalanobis (i.e., weighted Euclidean) distance metric over a real-vector-valued predictor space, in which small distances between two vectors imply similar target values. They note that this metric induces a kernel function whose parameters are set entirely from the data. Specifically, MLKR can learn a weight matrix W for a Mahalanobis metric that optimizes the leave-one out mean squared error of kernel regression (MSE), defined as:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2,$$

where $\hat{\mathbf{y}}_i$ is estimated using $\hat{\mathbf{y}}_j$ for all $i \neq j$, as in Eq. 1.

In MLKR, the weighted distance metric is learned using stochastic gradient descent. As an added benefit, MLKR is implicitly able to learn an appropriate kernel bandwidth.

3 String-Metric Learning for Kernel Regression (SMLKR)

Our novel contribution is an extension of MLKR to situations where the predictor variables are not real-valued vectors, but strings, and the distance metric we wish to learn is a weighted Levenshtein edit distance. Vector-valued representations of the strings themselves would only approximately preserve edit distance. Fortunately, it turns out that we do not need vector-valued representations of the strings at all. Define the *minimum edit-distance path* as the smallest-length sequence of edits that is needed to transform one string into another. Observe that the weighted edit distance between two strings s_i and s_j can be represented as the weighted sum of all the edits that must take place to transform one string into the other along the minimum edit-distance path (Bellet et al., 2012). In turn, these edits can be represented by a vector $\boldsymbol{\nu}_{ij}$ constructed as in Fig 1, while the weights can be represented by a vector $\mathbf{w} = (w_1, ..., w_M)^T$:

$$d_{WL}(s_i, s_j) = \sum_{m=1}^{M} w_m \nu_{ijm} = \mathbf{w}^T \boldsymbol{\nu}_{ij}.$$

Figure 1: Each element in $\boldsymbol{\nu}_{ij}$ (the vector at left) represents a type of edit. The entry ν_{ijm} represents the number of edits of type m that occur as string s_i (*boot*) is transformed into string s_j (*bee*).

Each entry of $\boldsymbol{\nu}_{ij}$ corresponds to a particular type of edit operation (e.g., substitution of character a for character b). The value assigned to each entry is the count of the total number of times that the corresponding edit operation must be applied to achieve transformation of string i to string j along the minimum edit-distance path.

We note that $\boldsymbol{\nu}_{ij}$ does not admit a unique representation, since there are multiple ways to transform one string to another in the same number of edits, using different edit operations. However, we adopt the convention that some class of edit operations always takes priority over another—e.g., that deletions always occur before substitutions. This then enables us to specify $\boldsymbol{\nu}_{ij}$ uniquely. We also adopt the convention that the weights for edit operations are symmetric—e.g., that the weight for substituting character a for character b is the same as the weight for substituting character b for character a, so we represent every such pair of edit operations by a single entry in $\boldsymbol{\nu}_{ij}$.

As in MLKR, our goal is to minimize the leave-one-out MSE,[1] where $k_{ij} = e^{-\mathbf{w}^T\boldsymbol{\nu}_{ij}}$. The gradient of the regression error for MSE is

$$\frac{\partial\mathcal{L}(\mathcal{D})}{\partial\mathbf{w}} = \frac{2}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\mathbf{y}}_i)\frac{\partial\hat{\mathbf{y}}_i}{\partial\mathbf{w}}$$

where

$$\frac{\partial\hat{\mathbf{y}}_i}{\partial\mathbf{w}} = \frac{\sum_{j\neq i}(\mathbf{y}_j - \hat{\mathbf{y}}_i)^T k_{ij}\boldsymbol{\nu}_{ij}}{\sum_{j\neq i}k_{ij}}.$$

Using this exact gradient, we can find the edit weights that minimize the loss function. We wish to constrain the weights to be non-negative, since weighted edit distance only makes sense with nonnegative weights. Thus, to minimize the loss we use the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm for box constraints (L-BFGS-B) (Byrd et al., 1995), a quasi-Newton method that allows bounded optimization. We made a Python implementation of SMLKR available at `http://bit.ly/25Hidqg/`.

[1] We attained similar results minimizing mean cosine error. The gradient for mean cosine error is

$$\frac{\partial\mathcal{L}(\mathcal{D})}{\partial\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N}\frac{(\|\hat{\mathbf{y}}_i\|\mathbf{y}_i - \mathcal{L}(\mathbf{y}_i,\hat{\mathbf{y}}_i)\hat{\mathbf{y}}_i)}{\|\hat{\mathbf{y}}_i\|^2}\frac{\partial\hat{\mathbf{y}}_i}{\partial\mathbf{w}}.$$

4 Experimental Setup

4.1 Data

Lexicon. A principal concern is the possibility that our models may detect morphemes rather than sub-morphemic units. To minimize this concern, we adopted an approach similar to that of Shillcock et al. (2001), of training our model only on monomorphemic words. Monomorphemic words were selected by cross-referencing the morphemic analyses contained in the CELEX lexical database (Baayen et al., 1996) with the morphemic analyses contained in the etymologies of the Oxford English Dictionary Online (`http://www.oed.com`). Then, we went through the filtered list and removed any remaining polymorphemic words as well as place names, demonyms, spelling variants, and proper nouns. Finally, words that were not among the 40,000 most frequent non-filler word types in the corpus were excluded. The final lexicon was composed of 4,949 word types.

Corpus and Semantic Model. The corpus we used to train our semantic model is a concatenation of the UKWaC, BNC, and Wikipedia corpora (Ferraresi et al., 2008; BNC Consortium, 2007; Parker et al., 2011). We trained our vector-space model on this corpus using the Word2Vec (Mikolov et al., 2013), as instantiated in the GENSIM package (Řehůřek and Sojka, 2010) for Python using default parameters. We produced 100-dimensional word-embedding vectors using the SkipGram algorithm of Word2Vec and normalized the 100-dimensional vector for each word so that its Euclidean norm was equal to 1.

4.2 Training

We trained SMLKR on the 100-dimensional Word2Vec embeddings using L-BGFS-B, and placing non-negativity constraints on the weights $\mathbf{w}$. We let SMLKR run until convergence, as de-

termined by the following criterion:

$$\frac{|\mathcal{L}^{(k-1)} - \mathcal{L}^{(k)}|}{\max(|\mathcal{L}^{(k-1)}|, |\mathcal{L}^{(k)}|)} = \epsilon$$

where $\mathcal{L}^{(k)}$ is the loss at the k^{th} iteration of learning, and we set $\epsilon = 2 \times 10^{-8}$. We randomly initialized the L-BGFS-B algorithm 10 times to avoid poor local minima, and kept the solution with the lowest loss.

5 Experiments

5.1 Model Analysis

Weighted Edit Distance Reveals More Non-Arbitrariness. We first assessed whether the structure found by kernel regression could arise merely by arbitrary, random pairings of form and meaning (i.e., strings and semantic vectors). We adopt a Monte Carlo testing procedure similar to the Mantel test of §2.1. We first randomly shuffled the assignment of the semantic vectors of all the words in the lexicon. We then trained SMLKR on the shuffled lexicon just as we did on the true lexicon. We measured the mean squared error of the SMLKR prediction. Out of 1000 reassignments, none produced a prediction error as small as the prediction error in the true lexicon (i.e., empirical p-value of $p < .001$).

For comparison, we analyzed our corpus using the correlation method of Monaghan et al. (2014b). In our implementation, we measured the correlation between the pairwise cosine distances produced by Word2Vec and pairwise orthographic edit distances for all pairs of words in our lexicon. The correlation between the Word2Vec semantic distances and the orthographic edit distances in our corpus was $r = 0.0194$, similar to the correlation reported by Monaghan et al. of $r = 0.016$ between the phoneme edit distances and the semantic distances in the monomorphemic English lexicon. We also looked at the correlation between the weighted edit distances produced by SMLKR and the Word2Vec semantic distances. The correlation between these distances was $r = 0.0464$; thus, the weighted edit distance captures more than 5.7 times as much variance as the unweighted edit distance. Further, using the estimated semantic vectors produced by the SMLKR model, we can actually produce new estimates of the semantic distances between the words. The correlation between these estimated semantic distances and the true semantic distances was $r = 0.1028$, revealing much more systematicity than revealed by the simple linear correlation method. The Mantel test with 1,000 permutations produced significant empirical p-values for all correlations ($p < .001$).

Systematicity Not Evenly Distributed Across Lexicon. What could be accounting for the higher degree of systematicity detected with SMLKR? Applying a more expressive model could result in a better fit simply because incidental but inconsequential patterns are being captured. Conversely, SMLKR could be finding phonosemantic sets which the correlation method of Monaghan et al. (Monaghan et al., 2014b) is unable to detect. We investigated further by determining what was driving the better fit produced by SMLKR. Monaghan et al. measure per-word systematicity as the change in the lexicon-wide form-meaning correlation that results from removing the word from the lexicon. The more the correlation decreases from removing the word, the more systematic the word is, according to this measure. They compared the distribution of this systematicity measure across the words in the lexicon to the distribution of systematicity in lexicons with randomly shuffled form-meaning assignments, and found that the null hypothesis that the distributions were identical could not be rejected. From this, they conclude that the observed systematicity of the lexicon is not a consequence only of small pockets of sound symbolism, but is rather a feature of the mappings from sound to meaning across the lexicon as a whole. However, it is possible that their methods may not be sensitive enough to find localized phonosemantic sets.

We developed our own measure of per-word systematicity by measuring the per-word regression error of the SMLKR model. We presume words with lower regression errors to be more systematic. A list of the words with the lowest per-word regression error in our corpus can be found in Table 1. Notably, many of these words, such as *fluff*, *flutter*, and *flick*, exhibit word beginnings or word endings that have been previously identified as phonaesthemes (Hutchins, 1998; Otis and Sagi, 2008). Others exhibit regular onomatopoeia, such as *clang* and *croak*.

We decided to investigate the distribution of systematicity across two-letter word-beginnings in our lexicon using a permutation test. The goal of the permutation test is to estimate a p-value for the

SMLKR	Correlation	Random
gurgle	emu	tunic
tingle	nexus	decay
hoop	asylum	skirmish
chink	ethic	scroll
swirl	odd	silk
ladle	slime	prom
flick	snare	knob
wobble	scarlet	havoc
tangle	deem	irate
knuckle	balustrade	veer
glitter	envoy	wear
twig	scrape	phone
fluff	essay	surgeon
rasp	ambit	hiccup
quill	echo	bowel
flutter	onus	sack
whirl	exam	lens
croak	pirouette	hovel
squeal	kohl	challenge
clang	chandelier	box

Table 1: *Left*: Most systematic words according to SMLKR. *Center*: Most systematic words according to the leave-one-out correlation method proposed by Monaghan et al. (2014b). *Right*: Randomly generated list for comparison.

likelihood that each set of words sharing a word beginning would exhibit the mean regression error it exhibits, if systematicity is randomly distributed across the lexicon. For each set $\mathcal{S}_\omega$ of words with word-beginning ω, we measured the mean SMLKR regression error of the words in $\mathcal{S}_\omega$. To get an empirical p-value for each $\mathcal{S}_\omega$ with cardinality greater than 5 (i.e., more than 5 word tokens), we randomly chose 10^5 sets of words in the lexicon with the same cardinality, and measured the mean SMLKR regression error for each of these random sets. If r of the randomly assembled sets had a lower mean regression error than $\mathcal{S}_\omega$ did, we assign an empirical p-value of $\frac{r}{10^5}$ to $\mathcal{S}_\omega$. A histogram of empirical p-values is in Fig. 2. From the figure, it seems clear that the p-values are not uniformly distributed; instead, an inordinate number of word-beginnings exhibit mean errors that are unlikely to occur if error is distributed arbitrarily across word-beginnings.

We can confirm this observation statistically. On the assumption that systematicity is arbitrarily distributed across word-beginnings, the empirical p-values of the permutation test should approximately conform to a Unif$(0, 1)$ distribution. We can test this hypothesis using a χ^2 test on the negative logarithms of the p-values (Fisher, 1932). Using this test, we reject the hypothesis that the p-values are uniformly distributed with $p < .0001$ ($\chi^2_{156} = 707.8$). The particular word-beginnings

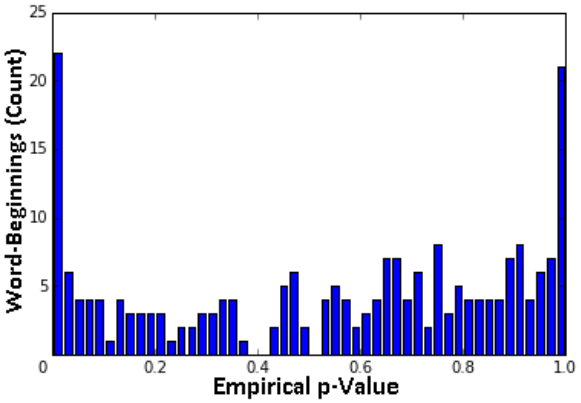

Figure 2: Histogram showing distribution of systematicity across two-letter word-beginnings, as measured by permutation-test empirical p-value.

Onset	p-value
fl-	$< 1 \times 10^{-4}$
sn-	$< 1 \times 10^{-4}$
sw-	$< 1 \times 10^{-4}$
tw-	$< 1 \times 10^{-4}$
gl-	1×10^{-3}
sl-	1×10^{-3}
bu-	1×10^{-3}
mu-	2×10^{-3}
wh-	2×10^{-3}
sc-/sk-	3×10^{-3}

Table 2: Word-beginnings with mean errors lower than predicted by random distribution of errors across lexicon. **Bold** are among the phonaesthemes identified by Hutchins (1998). *Italics* were identified by Otis and Sagi (2008).

with statistically significant empirical p-values ($p < .05$ after Benjamini-Hochberg (1995) correction for multiple comparisons) are in Table 2. Eight of these ten features are among the 18 two-letter onsets posited to be phonaesthemes by Hutchins (1998). For comparison, Otis and Sagi (2008) identified eight of Hutchins's 18 two-letter word-beginning candidate phonaesthemes (and 12 two-letter word-beginnings overall) as statistically significant, though they restricted their hypothesis space to only 50 pre-specified word-beginnings and word-endings. We are able to identify just as many candidate phonaesthemes, but with a much less restricted hypothesis space of candidates (225 rather than the 50 in Otis and Sagi's analysis) and with a general model not specifically attuned to finding phonaesthemes in particular, but rather systematicity in general.

5.2 Behavioral Evaluation of Systematicity Measure

We empirically tested whether the systematicity measure based on SMLKR regression error accords with naïve human judgments about how well-suited a word's form is to its meaning (its "phonosemantic feeling") (Stefanowitsch, 2002). We recruited 60 native English-speaking participants through Mechanical Turk, and asked them to judge the phonosemantic feeling of the 60 words in Table 1 on a sliding scale from 1 to 5.[2] We used Cronbach's α to measure inter-annotator reliability at $\alpha = 0.96$, indicating a high degree of inter-annotator reliability (Cronbach, 1951; George, 2000). The results showed that the words in the SMLKR list were rated higher for phonosemantic feeling than the words in the Correlation and Random lists. We fit a parametric linear mixed-effects model to the phonosemantic feeling judgments (Baayen et al., 2008), as implemented in the `lme4` library for R. As fixed effects, we entered the list identity (SMLKR, Correlation, Random), the word length, and the log frequency of the word in our corpus. Our random effects structure included a random intercept for word, and random subject slopes for all fixed effects, with all correlations allowed (a "maximal" random-effects structure (Barr et al., 2013)). Including list identity in the maximal mixed-effects model significantly improved model fit ($\chi^2_{11} = 126.08$, $p < 10^{-6}$). Post-hoc analysis revealed that the SMLKR list elicited average suitability judgments that were 0.49 points higher than the Random list ($p < 10^{-6}$) and 0.59 points higher than the Correlation list ($p < 10^{-6}$). Post-hoc analysis did not find a significant difference in suitability judgments between the Random and Correlation lists ($p > .16$).[3]

6 Conclusion

In this paper, we proposed SMLKR, a novel algorithm that can learn weighted string edit distances that minimize kernel regression error. We succeed in applying this algorithm to the problem of finding form-meaning systematicity in the monomorphemic English lexicon. Our algorithm offers improved global predictions of word-meaning given word-form at the lexicon-wide level. We show that this improvement seems related to localized pockets of form-meaning systematicity such as those previously uncovered in behavioral and corpus analyses. Unlike previous lexicon-wide analyses, we find that form-meaning systematicity is not randomly distributed throughout the English lexicon. Moreover, the measure of systematicity that we compute using SMLKR accords significantly with human raters' judgments about form-meaning correspondences in English.

Future work may investigate to what extent the SMLKR model can predict human intuitions about form-meaning systematicity in language. We do not know, for instance, if our model can predict human semantic judgments of novel words that have never been encountered. This is a question that has received attention in the market research literature, where new brand names are tested for the emotions they elicit (Klink, 2000). We would also like to investigate the degree to which our statistical model predicts the behavioral effects of phonosemantic systematicity during human semantic processing that have been reported in the psycholinguistics literature. Our model makes precise quantitative predictions that should allow us to address these questions.

While developing our model on preliminary versions of the monomorphemic lexicon, we noticed that the model detected high degrees of systematicity in words with suffixes such as *-ate* and *-tet* (e.g., *quintet*, *quartet*). We removed such words in the final analysis since they are polymorphemic, but this observation suggests that our algorithm may have applications in unsupervised morpheme discovery.

Finally, we would like to test our model using other representations of word-form and word-meaning. We chose to use orthographic rather than phonetic representations of words because of the variance in pronunciation present in the dialects of English that are manifested in our corpus. However, it would be interesting to verify our results in a phonological setting, perhaps using a monodialectal corpus. Moreover, previous local-level analyses suggest that systematicity seems to be concentrated in word-beginnings and word-

[2] Participants were given the following guidance: "Your job is to decide how well-suited each word is to what it means. This is known as the 'phonosemantic feeling.' Basically, most people feel like some of the words in their native language sound right, given what they mean." Full instructions and experiment available at `http://goo.gl/Z6Lzlp`

[3] Post hoc analyses were produced by comparing the items in only two of the lists at a time, and fitting the same mixed-effects model as above.

endings. Thus, it may be worthwhile to augment the representation of edit distance in our model by making it context-sensitive. Future work could also test whether a more interpretable meaning-space representation such as that provided by binary WordNet feature vectors reveals patterns of systematicity not found using a distributional semantic space.

Acknowledgments

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

References

Ekaterina Abramova, Raquel Fernández, and Federico Sangati. 2013. Automatic labeling of phonesthemic senses. In *Proeedings of the 35th Annual Conference of the Cognitive Science Society*, volume 35. Cognitive Science Society.

R Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1996. CELEX2 (CD-ROM).

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Todd M. Bailey and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4):568–591.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2012. Good edit similarity learning by loss minimization. *Machine Learning*, pages 5–35.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Benjamin K. Bergen. 2004. The psychological reality of phonaesthemes. *Language*, pages 290–311.

BNC Consortium. 2007. British National Corpus, Version 3 BNC XML edition.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Ferdinand de Saussure. 1916. *Course in General Linguistics*. McGraw-Hill, New York.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

John R. Firth. 1930. *Speech*. Benn's Sixpenny Library, London.

R.A. Fisher. 1932. *Statistical methods for research workers*. Oliver and Boyd, London.

Michael Gasser. 2004. The origins of arbitrariness in language. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, volume 26, pages 4–7.

Darren George. 2000. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update (4th ed.)*. Allyn & Bacon, London.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203:88–96.

Sharon Suzanne Hutchins. 1998. *The psychological reality, variability, and compositionality of English phonesthemes*. Ph.D. thesis, Emory University, Atlanta.

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

Willem J.M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01):1–38.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Margaret Magnus. 1998. *What's in a Word? Evidence for Phonosemantics*. Ph.D. thesis, University of Trondheim, Trondheim, Norway.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Padraic Monaghan, Gary Lupyan, and Morten H Christiansen. 2014a. The systematicity of the sign: Modeling activation of semantic attributes from nonwords. In P. Bello, M. Guarini, M. McShane, and B. Scassellati, editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 2741–2746, Austin, TX. Cognitive Science Society.

Padraic Monaghan, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. 2014b. How arbitrary is language? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1651).

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

Robert M. Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

John J. Ohala. 1984. An ethological perspective on common cross-language utilization of f0 of voice. *Phonetica*, 41(1):1–16.

Katya Otis and Eyal Sagi. 2008. Phonaesthemes: A corpus-based analysis. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 65–70.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. urlhttp://is.muni.cz/publication/884893/en.

Eyal Sagi and Katya Otis. 2008. Semantic glimmers: Phonaesthemes facilitate access to sentence meaning. In *9th Conference on Conceptual Structure, Discourse, & Language (CSDL9)*.

Richard Shillcock, Simon Kirby, Scott McDonald, and Chris Brew. 2001. Filled pauses and their status in the mental lexicon. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Anatol Stefanowitsch. 2002. Sound symbolism in a usage-driven model. Unpublished manuscript, Rice University, Houston, Texas, USA.

Monica Tamariz. 2006. *Exploring the adaptive structure of the mental lexicon*. Ph.D. thesis, University of Edinburgh, Edinburgh.

Monica Tamariz. 2008. Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2):259–278.

Killian Q. Weinberger and Gerald Tesauro. 2007. Metric learning for kernel regression. In *Eleventh International Conference on Artificial Intelligence and Statistics*, pages 608–615.