Meaning to Form: Measuring Systematicity as Information

Tiago Pimentel[♣] **Arya D. McCarthy**[♡] **Damián E. Blasi**[♠] **Brian Roark**[♢] **Ryan Cotterell**[♡†]
[♣]Kunumi, [♡]Johns Hopkins University, [♠]University of Zürich & MPI SHH,
[♢]Google, [†]University of Cambridge
tiago.pimentel@kunumi.com, arya@jhu.edu, damian.blasi@uzh.ch,
roarkbr@gmail.com, rdc42@cam.ac.uk

Abstract

A longstanding debate in semiotics centers on the relationship between linguistic signs and their corresponding semantics: is there an arbitrary relationship between a word form and its meaning, or does some systematic phenomenon pervade? For instance, does the character bigram *gl* have any systematic relationship to the meaning of words like *glisten*, *gleam* and *glow*? In this work, we offer a holistic quantification of the systematicity of the sign using mutual information and recurrent neural networks. We employ these in a data-driven and massively multilingual approach to the question, examining 106 languages. We find a statistically significant reduction in entropy when modeling a word form conditioned on its semantic representation. Encouragingly, we also recover well-attested English examples of systematic affixes. We conclude with the meta-point: Our approximate effect size (measured in bits) is quite small—despite some amount of systematicity between form and meaning, an arbitrary relationship and its resulting benefits dominate human language.

1 Introduction

Saussure (1916) expounded on the **arbitrariness of the sign**. Seen as a critical facet of human language (Hockett, 1960), the idea posits that a sign in human language (a word, in our inquiry) is structured at two levels: the **signified**, which captures its meaning, and the **signifier**, which has no meaning but manifests the form of the sign. Saussure himself, however, also documented instances of sound symbolism in language (Saussure, 1912). In this paper, we present computational evidence of relevance to both aspects of Saussure's work.

While dominant among linguists, arbitrariness has been subject to both long theoretical debate (Wilkins, 1668; Eco, 1995; Johnson, 2004; Pullum and Scholz, 2007) and numerous empirical and experimental studies (Hutchins, 1998; Bergen, 2004; Monaghan et al., 2011; Abramova and Fernández, 2016; Blasi et al., 2016; Gutierrez et al., 2016; Dautriche et al., 2017). Taken as a whole, these studies suggest non-trivial interactions in the form–meaning interface between the signified and the signifier (Dingemanse et al., 2015).

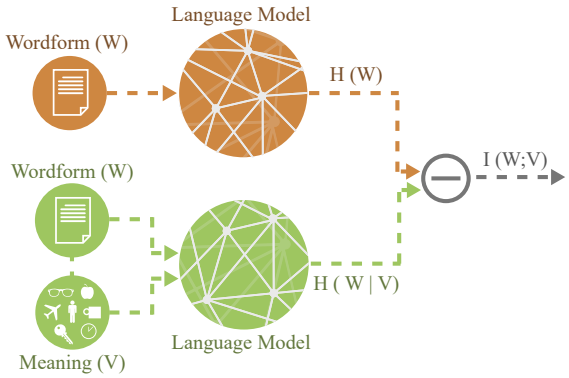

Figure 1: We use two independent language models to estimate the mutual information between word forms and meaning—i.e. systematicity, as per our definition. The language models provide upper bounds on $\mathrm{H}(W)$ and $\mathrm{H}(W \mid V)$, which can be used to estimate $\mathrm{I}(W;V)$. This estimate is as good as the upper bounds are tight—see discussion in §3.4.

Although the new wave of studies on form–meaning associations range across multiple languages, methods and working hypotheses, they all converge on two important dimensions:

1. The description of meaning is parameterized with pre-defined labels—e.g., by using existing ontologies like List et al. (2016).
2. The description of forms is restricted to the presence, absence or sheer number of occurrence of particular units (such as phones, syllables or handshapes).

We take an information-theoretic approach to quan-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1751–1764
Florence, Italy, July 28 - August 2, 2019.

tifying the relationship between form and meaning using flexible representations in both domains, rephrasing the question of systematicity: *How much does certainty of one reduce uncertainty of the other?* This gives an operationalization as the **mutual information** between form and meaning, when treating both as random variables—the signifier as a word's phone string representation in the International Phonetic Alphabet (IPA), and the signified as a distributed representation (Mikolov et al., 2013) for that word's lexical semantics, devoid of morphological or other subword information. We show how to estimate mutual information as the difference in entropy of two phone-level LSTM language models—one of which is conditioned on the semantic representation. This operationalization, depicted in Figure 1, allows us to express the *global* effect of meaning on form in vocabulary datasets with wide semantic coverage.

In addition to this lexicon-level characterization of systematicity, we also show that this paradigm can be leveraged for studying more narrowly-defined form-meaning associations such as **phonesthemes**—submorphemic, meaning-bearing units—in the style of Gutierrez et al. (2016). These short sound sequences typically suggest some aspect of meaning in the words that contain them, like *-ump* for rounded things in English. Previous computational studies, whether focusing on characterizing the degree of systematicity (Monaghan et al., 2014b,a, 2011; Shillcock et al., 2001), discovering phonesthemes (Liu et al., 2018), or both (Gutierrez et al., 2016), have invariably framed systematicity in terms of distances and/or similarities–the relation between word-form distance/similarity on the one hand (e.g., based on string edit distance) and semantic distance/similarity on the other (e.g., as defined within a semantic vector space). Our methods have the virtue of not relying on some predefined notion of similarity or distance in either domain for our measurement of systematicity.

Empirically, we focus on two experimental regimes. First, we focus on a large corpus (CELEX) of phone transcriptions in Dutch, English, and German. In these three languages, we find a significant yet small mutual information even when controlling for grammatical category. Second, we perform a massively multilingual exploration of sound–meaning systematicity (§5.1) on the NorthEuraLex corpus (Dellert and Jäger, 2017). This corpus contains expanded Swadesh lists in 106 languages using a unified alphabet of phones. It contains 1016 words in each language, which is often not enough to detect systematicity—we trade the coverage of CELEX for the breadth of languages. Nevertheless, using our information-theoretic operationalization, in most of the languages considered (87 of 106), we find a statistically significant reduction in entropy of phone language modeling by conditioning on a word's meaning (§5.2). Finally, we find a weak positive correlation between our computed mutual information and human judgments of form–meaning relatedness.

2 Systematic form-meaning associations

2.1 Arbitrariness

The lack of a forceful association between form and meaning is regarded as a design feature of language (Hockett, 1960). This arbitrariness of the sign is thought to provide a flexible and efficient way for encoding new referents (Monaghan et al., 2011). It has been claimed that it enhances learnability because newly acquired concepts can be paired to any word, instead of devising the word that properly places the concept in one's constellation of concepts (Gasser et al., 2005), and that it facilitates mental processing compared to an icon-based symbol system, in that the word–meaning map can be direct (Lupyan and Thompson-Schill, 2012). Most importantly, decoupling form from meaning allows communication about things that are not directly grounded in percepts (Clark, 1998; Dingemanse et al., 2015). This opens the door to another of Hockett (1960)'s design features of language: duality of patterning (Martinet, 1949), the idea that language exists on the level of meaningless units (the *distinctive*; typically phonemes) composed to form the level of meaningful units (the *significant*; typically morphemes).

2.2 Non-arbitrariness and systematicty

Contemporary research has established that non-arbitrary form-meaning associations in vocabulary are more common and diverse than previously thought (Dingemanse et al., 2015). Some non-arbitrary associations might be found repeatedly across unrelated languages presumably due to species-wide cognitive biases (Blasi et al., 2016), others are restricted to language-specific word classes that allow for more or less transparent iconic mappings – so-called *ideophones*, see Dingemanse (2012; 2018) – and yet others might emerge

from properties of discourse and usage rather than meaning per se (Piantadosi et al., 2011).

Systematicity is meant to cover all cases of non-arbitrary form-meaning associations of moderate to large presence in a vocabulary within a language (Dingemanse et al., 2015). In morphology-rich languages, systematic patterns are readily apparent: for instance, across a large number of languages recurring TAM markers or transitivity morphemes could be used to detect verbs, whereas case markers or nominalizing morphemes can serve as a cue for nouns. Yet a sizable portion of research on systematicity is geared towards subtle patterns at the word root level, beyond any ostensive rules of grammar.

By and large, systematicity is hailed as a trait easing language acquisition. It reduces the radical uncertainty humans find when first encountering a new word by providing clues about category and meaning (Monaghan et al., 2014a). Systematic patterns can display a large scope within a language: for instance, systematic associations distinguishing nouns from verbs have been found in every language where a comparison was performed systematically (e.g. Monaghan et al., 2007). But at its extreme, systematicity would manifest as an ontology encoded phonetically, e.g., all plants begin with the letter 'g', and animals with the letter 'z' (Wilkins, 1668; Eco, 1995). As Dingemanse et al. (2015) note, a system of similar forms expressing similar meanings "would lead to high confusability of the very items most in need of differentiation".

2.3 Phonesthemes

One particular systematic pattern comes in the form of **phonesthemes** (Firth, 1964). These are submorphemic and mostly unproductive affixal units, usually flagging a relatively small semantic domain. A classic example in English is *gl-*, a prefix for words relating to light or vision, e.g. *glimmer*, *glisten*, *glitter*, *gleam*, *glow* and *glint* (Bergen, 2004).

Phonesthemes have psychological import; they can be shown to accelerate reaction times in language processing (Hutchins, 1998; Bergen, 2004; Magnus, 2000). They have been attested in English (Wallis, 1699; Firth, 1930; Marchand, 1959; Bolinger, 1949, 2014), Swedish (Abelin, 1999), Japanese (Hamano, 1998), Ojibwa (Rhodes, 1981), Hmong (Ratliff, 1992), and myriad Austronesian languages (McCune, 1985; Blust, 1988). In fact, as Bergen (2004) notes, "every systematic study of a particular language has produced results suggesting that that language has phonesthemes". Liu et al. (2018) survey computational approaches for identifying phonesthemes.

3 Estimating Systematicity with Information Theory

3.1 Notation and formalization

Following Shillcock et al. (2001), we define a sign as a tuple $(\mathbf{v}^{(i)}, \mathbf{w}^{(i)})$ of a word's distributional semantic representation (a vector) and its phone string representation (a word form). For a natural language with a set of phones Σ (including a special end-of-string token), we take the space of word forms to be Σ^*, with $\mathbf{w}^{(i)} \in \Sigma^*$. We treat the semantic space as a high-dimensional real vector space $\mathbb{R}^d$, with $\mathbf{v}^{(i)} \in \mathbb{R}^d$. The particular $\mathbf{v}^{(i)}$ and $\mathbf{w}^{(i)}$ are instances of random variables V and W.

Further, we want to hunt down potential phonesthemes; we define these to be phone sequences which, compared to others of their length, have a larger mutual information with their meaning. We eliminate positional confounds by examining only words' prefixes $\mathbf{w}_{<k}$ and suffixes $\mathbf{w}_{>k}$.[1]

3.2 A variational upper bound

Entropy, the workhorse of information theory, captures the uncertainty of a probability distribution. In our language modeling case, the quantity is

$$\mathrm{H}(W) \equiv \sum_{\mathbf{w}\in\Sigma^*} \Pr(\mathbf{w}) \log \frac{1}{\Pr(\mathbf{w})}. \tag{1}$$

Entropy is the average number of bits required to represent a string in the distribution, under an optimal coding scheme. When computing it, we are faced with two problems: We do not know the distribution over word-forms $\Pr(W)$ and, even if we did, computing Equation 1 requires summing over the infinite set of possible strings Σ^*.

We follow Brown et al. (1992) in tackling these problems together. Approximating $\Pr(W)$ with any known distribution $Q(W)$, we get a variational upper bound on $\mathrm{H}(W)$ from their cross-entropy, i.e.

$$\mathrm{H}(W) \leq \mathrm{H}_Q(W) \tag{2a}$$

$$= \sum_{\mathbf{w}\in\Sigma^*} \Pr(\mathbf{w}) \log \frac{1}{Q(\mathbf{w})}. \tag{2b}$$

[1] In line with, e.g., Cucerzan and Yarowsky (2003), we treat affixes as word-initial or word-final sequences, regardless of their status as attested morphological entities.

Equation 2b still requires knowledge of $\Pr(W)$ and involves an infinite sum, though. Nonetheless, we can use a finite set $\tilde{\mathbf{W}}$ of samples from $\Pr(W)$ to get an empirical estimate of this value.

$$\mathrm{H}_Q(W) \approx \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{Q\left(\tilde{\mathbf{w}}^{(i)}\right)}, \quad \tilde{\mathbf{w}}^{(i)} \in \tilde{\mathbf{W}} \sim \Pr(W) \tag{3}$$

with equality if we let $N \to \infty$.[2] We now use Equation 3 as an estimate for the entropy of a lexicon.

Conditional entropy Conditional entropy reflects the average *additional* number of bits needed to represent an event, given knowledge of another random variable. If V completely determines W, then the quantity is 0. Conversely, if the variables are independent, then $\mathrm{H}(W) = \mathrm{H}(W \mid V)$. Analogously to the unconditional case, we can get an upper bound for the conditional entropy by approximating $\Pr(W \mid V)$ with another distribution Q.

$$\mathrm{H}_Q(W \mid V) \approx \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{Q\left(\tilde{\mathbf{w}}^{(i)} \mid \tilde{\mathbf{v}}^{(i)}\right)} \tag{4}$$

where $(\tilde{\mathbf{w}}^{(i)}, \tilde{\mathbf{v}}^{(i)}) \sim \Pr(W,V)$.

3.3 Systematicity as mutual information

Mutual information (I) measures the amount of information (bits) that the knowledge of either form or meaning provides about the other. It is the difference between the entropy and conditional entropy:

$$\begin{aligned} \mathrm{I}(W;V) &\equiv \mathrm{H}(W) - \mathrm{H}(W \mid V) && \text{(5a)} \\ &\approx \mathrm{H}_Q(W) - \mathrm{H}_Q(W \mid V). && \text{(5b)} \end{aligned}$$

Systematicity will thus be framed as (statistically significant) nonzero mutual information $\mathrm{I}(V;W)$.

3.4 Learning Q

Our method relies on decomposing mutual information into a difference of entropies, as shown in Equation 5b. We use upper bounds on both the entropy and conditional entropy measures, so our calculated mutual information is an estimate.

This estimate is as good as our bounds are tight, being perfect when $\Pr(W) = Q(W)$ and $\Pr(W|V) = Q(W|V)$. Still, as we subtract two upper bounds, we cannot guarantee that our MI estimate approaches the real MI from above or below because we do not know which of the entropies' bounds are tighter. There is nothing *principled* that we can say about the result, except that it is consistent.

The procedure for learning the distribution Q is, thus, essential to our method. We must first define a family of distributions Ψ from which Q is learned. Then, we learn $Q \in \Psi$ by minimizing the right-hand-size of Equation 2b—which corresponds to maximum likelihood estimation

$$Q = \arg\inf_{q \in \Psi} \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{q\left(\tilde{\mathbf{w}}^{(i)}\right)}. \tag{6}$$

In this work, we employ a state-of-the-art phone-level LSTM language model as our Ψ to approximate $\Pr(W)$ as closely as possible.

3.5 Recurrent neural LM

A phone-level language model (LM) provides a probability distribution over Σ^*:

$$\Pr(\mathbf{w}) = \prod_{i=1}^{|\mathbf{w}|+1} \Pr\left(w_i \mid \mathbf{w}_{<i}\right). \tag{7}$$

Recurrent neural networks are great representation extractors, being able to model long dependencies—up to a few hundred tokens (Khandelwal et al., 2018)—and complex distributions $\Pr(w_i \mid \mathbf{w}_{<i})$ (Mikolov et al., 2010; Sundermeyer et al., 2012). We choose LSTM language models in particular, the state-of-the-art for character-level language modeling (Merity et al., 2018).[3]

Our architecture embeds a word—a sequence of tokens $w_i \in \Sigma$—using an embedding lookup table, resulting in vectors $\mathbf{z}_i \in \mathbb{R}^d$. These are fed into an LSTM, which produces high-dimensional representations of the sequence (hidden states):

$$\mathbf{h}_j = \mathrm{LSTM}\left(\mathbf{h}_{j-1}, \mathbf{z}_j\right), \quad j \in \{1, \ldots, n+1\}, \tag{8}$$

where $\mathbf{h}_0$ is the zero vector. Each hidden state is linearly transformed and fed into a softmax function, producing a distribution over the next phone: $\Pr\left(w_i \mid \mathbf{w}_{<i}\right) = \mathrm{softmax}\left(\mathbf{W}\mathbf{h}_i + \mathbf{b}\right)$.

4 Experimental Design

4.1 Datasets

We first analyze the CELEX database (Baayen et al., 1995), which provides many word types for Dutch, English, and German. In measuring systematicity, we control for morphological variation by only considering monomorphemic words, as in

[2] This is a direct consequence of the law of large numbers.

[3] Our tokens are phones rather than graphemes.

Dautriche et al. (2017). Our type-level resource contains lemmata, eliminating the noisy effect of morphologically inflected forms. CELEX contains 6040 English, 3864 German, and 3603 Dutch lemmata for which we have embeddings.

While CELEX is a large, well annotated corpus, it only spans three lexically related languages. The NorthEuraLex database (Dellert and Jäger, 2017) is thus appealing. It is a lexicon of 1016 "basic" concepts, written in a unified IPA scheme and aligned across 107 languages that span 21 language families (including isolates).[4] While we cannot restrict NorthEuraLex to monomorphemic words (because it was not annotated by linguists and segmentation models are weak for its low-resource languages), it mainly contains word types for basic concepts—e.g., animal names or verbs—so we are comfortable in the modeling assumption that the words are not decomposable into multiple morphemes.

Unlike Dautriche et al. (2017), who draw lexicons from Wikipedia, or Otis and Sagi (2008), we directly use a phone string representation, rather than their proxy of using each language's orthography. This makes our work the first to quantify the interface between phones and meaning in a massively multilingual setting.

Blasi et al. (2016) is the only large-scale exploration of phonetic representations that we find. They examine 40 aligned concepts over 4000 languages and identify that sound correspondences exist across the vast majority. Their resource (Wichmann et al., 2018) does not have enough examples to train our language models, and we add to their findings by measuring a relationship *between form and meaning*, rather than form given meaning.

4.2 Embeddings

We use pre-trained WORD2VEC representations as meaning vectors for the basic concepts. For CELEX, specific representations were pre-trained for each of the three languages.[5] For NorthEuraLex, as its words are concept aligned, we use the same English vectors for all languages. Pragmatically, we choose English because its vectors have the largest coverage of the lexicon. This does not mean that we assume that semantic spaces across languages to be strictly comparable. In fact, we would expect that more direct methods of estimating these vectors would be preferable if they were practical.

Note that the methods described above are likely underestimating the semantic systematicity in the data, for a couple of reasons. First, WORD2VEC and other related methods have been shown to do a better job at capturing general relatedness rather than semantic similarity per se (Hill et al., 2015). Second, our use of the English vectors across the concept-aligned corpora is a somewhat coarse expedient. To the extent that the English serves as a poor model for the other languages, we should expect smaller MI estimates. In short, we have chosen easy-to-replicate methods based on commonly used models, rather than extensively tuning our approach for these experiments, possibly at the expense of the size of the effect we observe.

To reduce spurious fitting to noise in the dataset, we reduce the dimensionality of these vectors from the original 300 to d while capturing maximal variance, using principal components analysis (PCA).

These resulting d-dimensional vectors are kept fixed while training the conditional language model. Each d-dimensional vector $\mathbf{v}$ is linearly transformed to serve as the initial hidden state of the conditional LSTM language model:

$$\mathbf{h}_0 = \mathbf{W}^{(v)}\mathbf{v} + \mathbf{b}^{(v)}$$
$$\mathbf{h}_j = \text{LSTM}(\mathbf{h}_{j-1}, \mathbf{z}_j), \quad j \in \{1, \ldots, n+1\}.$$

We reject morphologically informed embeddings (e.g., Bojanowski et al., 2017) because this would be circular: We cannot question the arbitrariness of the form–meaning interface if the meaning representations are constructed with explicit information from the form. This is the same reason that we do not fine-tune the embeddings—our goal is to enforce as clean a separation as possible of model and form, then suss out what is inextricable.

4.3 Controlling for grammatical category

The value of WORD2VEC comes from distilling more than just meaning. It also encodes the grammatical classes of words. Unfortunately, this is a trivial source of systematicity: if a language's lemmata for some class follow a regular pattern (such as the verbal infinitive endings in Romance languages), our model will have uncovered something meaningless. Prior work—e.g., (Dautriche et al., 2017; Gutierrez et al., 2016)—does not account for

[4] We omit Mandarin; the absence of tone annotations leaves its phonotactics greatly underspecified. All reported results are for the remaining 106 languages.

[5] We use Google's WORD2VEC representations pre-trained in Google News corpus for English, while WORD2VEC was trained using Wikipedia dumps for German and Dutch with default hyper-parameters.

this. To isolate factors like these, we can estimate the mutual information between word form and meaning, while conditioning on a third factor. The expression is similar to Equation 5a:

$$\mathrm{I}(W;V \mid C) \equiv \mathrm{H}(W \mid C) - \mathrm{H}(W \mid V,C), \qquad (9)$$

where C is our third factor—in this case, grammatical class.[6]

Both CELEX and NorthEuraLex are annotated with grammatical classes for each word. We create a lookup embedding for each class in a language, then use the resulting representation as an initial hidden state to the LSTM ($\mathbf{h}_0 = \mathbf{c}$). When conditioning on both meaning and class, we concatenate half-sized representations of the meaning (pre-trained) and class to create the first hidden state ($\mathbf{h}_0 = [\mathbf{c}'; \mathbf{W}^{(v)}\mathbf{v}' + \mathbf{b}^{(v)}]$).

4.4 Hypothesis testing

We follow Gutierrez et al. (2016) and Liu et al. (2018) in using a permutation test to assess our statistical significance. In it, we randomly swap the sign of I values for each word, showing mutual information is significantly positive. Our null hypothesis, then, is that this value should be 0. Recomputing the average mutual information over many shufflings gives rise to an empirical p-value: asymptotically, it will be twice the fraction of permutations with a higher mutual information than the true lexicon. In our case, we used 100,000 random permutations.

4.5 Hyperparameters and optimization

We split both datasets into ten folds, using one fold for validation, another for testing, and the rest for training. We optimize all hyper-parameters with 50 rounds of Bayesian optimization—this includes the number of layers in the LSTM, its hidden size, the PCA size d used to compress the meaning vectors, and a dropout probability. Such an optimization is important to get tighter bounds for the entropies, as discussed in §3.4. We use a Gaussian process prior and maximize the expected improvement on the validation set, as in Snoek et al. (2012). [7]

5 Results and Analysis

5.1 Identifying systematicity

We find statistically significant nonzero mutual information in all three CELEX languages (Dutch, English, and German), using a permutation test to establish significance. This gives us grounds to reject the null hypothesis. We also find a statistically significant mutual information when conditioning entropies in words' grammar classes. These results are summarized in Table 1.

But how much *could* the mutual information have been? A raw number of bits is not easily interpretable, so we provide another information-theoretic quantity, the **uncertainty coefficient**, expressing the fraction of bits we can predict given the meaning: $\mathrm{U}(W \mid V) = \frac{\mathrm{I}(W;V)}{\mathrm{H}(W)}$.The mutual information $\mathrm{I}(W;V)$ is upper-bounded by the language's entropy $\mathrm{H}(W)$, so the uncertainty coefficient is between zero and one.[8] For the CELEX data, we give the uncertainty coefficients with and without conditioning on part of speech in Table 1.

By comparing results with and without conditioning on grammatical category, we see the importance of controlling for known factors of systematicity. As expected, all systematicity (mutual information) results are smaller when we condition on part of speech. After conditioning, systematicity remains present, though. In English, we can guess about 3.25% of the bits encoding the phone sequence, given the meaning. In Dutch and German, these quantities are higher. The effect size of systematicity in these languages, though, is small.

5.2 Broadly multilingual analysis

On the larger set of languages in NorthEuraLex, we see that 87 of the 106 languages have statistically significant systematicity ($p < 0.05$), after Benjamini–Hochberg (1995) corrections. When we control for grammatical classes ($\mathrm{I}(W;V \mid \mathrm{POS})$), we still get significant systematicity across languages ($p < 10^{-3}$). A per-language analysis, though, only finds statistical significance for 17 of them, after Benjamini–Hochberg (1995) corrections. This evinces the importance of conditioning on grammatical category; without doing so, we would find a spurious result due to crafted, mor-

[6] If markers of subclasses within a given part of speech are frequent, these may also emerge.

[7] Our implementation is available at https://github.com/tpimentelms/meaning2form.

[8] Because of our estimation, it may be less than zero.

		Systematicity			Systematicity controlling for POS tags		
Language	H(W)	I($W;V$)	U($W \mid V$)	Cohen's d	I($W;V \mid$ POS)	U($W \mid V;$POS)	Cohen's d
English	3.401	0.110	3.24%	0.175	0.084	2.50%	0.133
German	3.195	0.168	5.26%	0.221	0.154	4.84%	0.203
Dutch	3.245	0.156	4.82%	0.222	0.089	2.84%	0.123

Table 1: Mutual information (in bits per phone), uncertainty coefficients, and Cohen's effect size results for CELEX. Per-phone word–form entropy added for comparison. All mutual information values are statistically significant ($p < 10^{-5}$), as tested with a permutation test with 10^5 permutations.

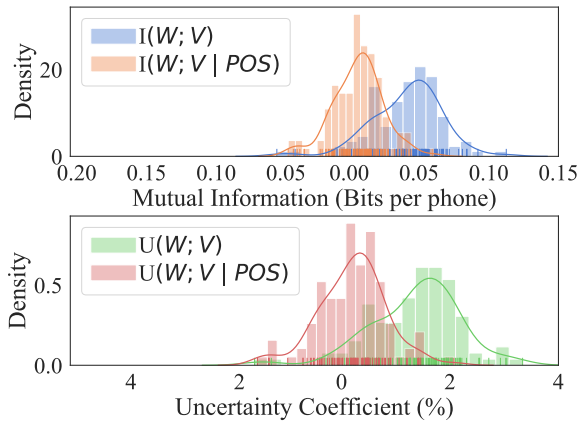

Figure 2: Mutual information and uncertainty coefficients for each language of NorthEuraLex.

phological systematicity. We present kernel density estimates for these results in Figure 2 and give full results in Appendix A. Across all languages, the average uncertainty coefficient was 1.37% (Cohen's d 0.1936). When controlling for grammatical classes, though, it was only 0.2% (Cohen's d 0.0287).

There were only 970 concepts with corresponding WORD2VEC representations in this dataset, and our language models easily overfit when conditioned on these. As we optimize the used number of PCA components (d) for these word embeddings, we can check its 'optimum' size. The average d across NorthEuraLex languages was only ≈ 22, while on CELEX it was ≈ 153. This might imply that the model couldn't find systematicity in some languages due to the dataset's small size—models were too prone to overfitting.

5.3 Fantastic phonesthemes and where to find them

As a phonestheme is, by definition, a sequence of phones that suggest a particular meaning, we expect them to have higher mutual information values when compared to other k-grams in the lexicon—measured in bits per phone. To identify that a prefix of length k, $\mathbf{w}_{\leq k}$, is a phonestheme, we compare it to all such prefixes, being interested in the mutual information $\mathrm{I}(W_{\leq k}, V)$. For each prefix in our dataset, we compute the average mutual information over all n words it appears in.We then sample 10^5 other sets of n words and get their average mutual information. Each prefix is identified as a phonestheme with a p-value of $\frac{r}{10^5}$, where r is how many comparison where it has a lower systematicity than the random sets.[9] Table 2 shows identified phonesthemes for English, Dutch, and German.

Inspecting the German data, it is clear that some of these prefixes and affixes that we find are fossilized pieces of derivational etymology. Further, many of the endings in German are simply the verb ending -/ən/ with an additional preceding phone. Dutch and English are less patterned. While we find few examples in Dutch, all are extremely significant. It can be argued that two examples (-/əl/ and -/xt/) are not semantic markers but rather categorizing heads in the framework of distributed morphology (Marantz and Halle, 1993)—suggestions that the words are nouns. Further, in English, we find other examples of fossilized morphology, (/kən/-) and (/ɪn/-). In this sense, our found phonesthemes are related to another class of restricted-application subword: bound morphemes (Bloomfield, 1933; Aronoff, 1976; Spencer, 1991), which carry known meaning and cannot occur alone.

From the list of English prefix phonesthemes we present here, all but /ɪn/- and /kən/- find support in the literature (Hutchins, 1998; Otis and Sagi, 2008; Gutierrez et al., 2016; Liu et al., 2018). Furthermore, an interesting case is the suffix -/mp/, which is identified with a high confidence. This might be picking up on phonesthemes -/ump/ and -/amp/

[9] While this explanation is specific to prefixes, we straightforwardly applied this to suffixes by reversing the word forms—e.g. '*banana*' ↦ '*ananab*'.

Language	Phonestheme	Count	Examples	p-value
Dutch	/sx/-	110	schelp, schild, schot, shacht, schaar	<0.00001
	-/əl/	124	kegel, nevel, beitel, vleugel, zetel	<0.00001
	-/xt/	42	beicht, nacht, vocht, plicht, licht	<0.00001
	-/ɔp/	21	stop, shop, drop, top, bob	0.00068
English	/ɪn/-	33	infidel, intellect, institute, enigma, interim	<0.00001
	/sl/-	59	slop, slough, sluice, slim, slush	<0.00001
	-/kt/	36	aspect, object, fact, viaduct, tact	0.00001
	-/mə/	32	panorama, asthma, trachoma, eczema, magma	0.00002
	-/mp/	44	stump, cramp, pump, clamp, lump	0.00003
	-/əm/	62	millennium, amalgam, paroxysm, pogrom, jetsam	0.00007
	/fl/-	64	flaw, flake, fluff, flail, flash	0.00009
	/bʊ/-	35	bum, bunch, bunk, butt, buck	0.00013
	-/ʔp/	23	hop, strop, plop, pop, flop	0.00032
	/gl/-	28	gleam, gloom, glaze, glee, glum	0.00046
	/sn/-	38	sneak, snide, snaffle, snout, snook	0.00077
	-/nə/	34	henna, savanna, fauna, alumna, angina	0.00102
	-/æg/	23	swag, shag, bag, mag, gag	0.00107
	/sw/-	43	swamp, swoon, swish, swoop, swig	0.00112
	/sɪ/-	78	silica, secede, silicone, secrete, cereal	0.00198
	-/kə/	22	japonica, yucca, mica, hookah, circa	0.00217
	/sɛ/-	34	shell, sheriff, shelf, chevron, shed	0.00217
	/kən/-	31	conceal, condemn, concert, construe, continue	0.00429
German	/gə/-	69	geschehen, Gebiet, gering, Geruecht, gesinnt	<0.00001
	-/əln/	58	rascheln, rumpeln, tummeln, torkeln, mogeln	<0.00001
	-/ln/	58	rascheln, rumpeln, tummeln, torkeln, mogeln	<0.00001
	-/ən/	801	goennen, saeen, besuchen, giessen, streiten	<0.00001
	/ɪn/-	34	Indiz, indes, intern, innehaben, innerhalb	<0.00001
	/bə/-	32	bestaetigen, beweisen, bewerkstelligen, betrachten, beschwichtigen	<0.00001
	-/pə/	36	Lampe, Klappe, Kappe, Raupe, Wespe	0.00002
	-/ʃən/	24	dreschen, wischen, mischen, rauschen, lutschen	0.00002
	/ʃl/-	39	schlagen, schlingen, schleifen, schleudern, schluepfen	0.00015
	-/kən/	76	backen, strecken, spucken, druecken, schmecken	0.00016
	-/tsən/	47	blitzen, schwatzen, duzen, stanzen, einschmelzen	0.00026
	-/lən/	41	quellen, prellen, johlen, bruellen, eilen	0.00029
	/ain/-	25	einstehen, eintreiben, einmuenden, einfinden, eingedenk	0.00033
	-/ɪx/	59	reich, weich, bleich, gleich, Laich	0.00033
	/ʃn/-	22	schnitzen, schnalzen, schnappen, schnurren, schneiden	0.00036
	/ʃm/-	23	schmieren, schmieden, schmunzeln, schmoren, schmeissen	0.00077
	/ʃv/-	38	schweben, schweifen, schwirren, schwellen, schwimmen	0.00124
	-/rən/	62	servieren, wehren, sparen, kapieren, hantieren	0.00247
	/br/-	35	brausen, bremsen, brechen, brennen, brauen	0.00258
	-/tə/	86	Paste, Quote, Kette, vierte, Sorte	0.00281
	-/nə/	66	Traene, Tonne, Laterne, Fahne, Spinne	0.00354
	-/ərn/	70	schillern, schimmern, kapern, knattern, rattern	0.00365

Table 2: Discovered phonesthemes, represented as IPA, in Dutch, English, and German, sorted p-values according to the Benjamini–Hochberg (1995) correction. Count refers to the number of types in our corpus with that affix.

from Hutchins (1998)'s list.

5.4 Correlation with human judgments

As a final, albeit weak, validation of our model, we consider how well our computed systematicity compares to human judgments (Hutchins, 1998; Gutierrez et al., 2016; Liu et al., 2018). We turn to the survey data of Liu et al. (2018), in which workers on Amazon Mechanical Turk gave a 1-to-5 judgment of how well a word's form suited its meaning. For each of their model's top 15 predicted phonesthemes and 15 random non-predicted phonesthemes, the authors chose five words containing the prefix for workers to evaluate.[10] Comparing these judgments to our model-computed estimates of mutual information $\mathrm{I}(W_{<2};V)$, we find a weak, positive Spearman's rank correlation ($\rho = 0.352$ with $p = 0.03$). This shows that prefixes for which we find higher systematicity—according to mutual information—also tend to have higher human-judged systematicity.

6 Conclusion

We have revisited the linguistic question of the arbitrariness—and the systematicity—of the sign. We have framed the question on information-theoretic grounds, estimating entropies by state-of-the-art neural language modeling. We find evidence in 87 of 106 languages for a significant systematic pattern between form and meaning, reducing approximately 5% of the phone-sequence uncertainty of German lexicons and 2.5% in English and Dutch, when controlling for part of speech.

We have identified meaningful phonesthemes according to our operationalization, and we have good precision—all but two of our English phonesthemes are attested in prior work. An avenue for future work is connecting our discovered phonesthemes to putative meanings, as done by Abramova et al. (2013) and Abramova and Fernández (2016).

The low uncertainty reduction suggests that the lexicon is still largely arbitrary. According to the information-theoretic perspective of Monaghan et al. (2011), an optimal lexicon has an arbitrary mapping between form and meaning. If this is true, then a large amount of these benefits do accrue to language; that is, given the small degree of systematicity, we lose little of the benefit.

[10] Of the 150 judgements in their dataset, only 35 were in ours as well, so we restrict our analysis to them. This is a weak signal for our model's validity.

Acknowledgments

The authors would like to thank Mark Dingemanse, Adina Williams, and the anonymous reviewers for valuable insights and useful suggestions.

References

Åsa Abelin. 1999. *Studies in sound symbolism*. Ph.D. thesis, Department of Linguistics, Göteborg University Göteborg.

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

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

Mark Aronoff. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.*, 1:1–134.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX2 lexical database (release 2); LDC96L14. *Distributed by the Linguistic Data Consortium, University of Pennsylvania, web download*.

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)*, 57(1):289–300.

Benjamin K Bergen. 2004. The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston.

Robert A Blust. 1988. *Austronesian root theory: An essay on the limits of morphology*, volume 19. John Benjamins Publishing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Dwight Bolinger. 2014. *Language-the loaded weapon: The use and abuse of language today*. Routledge.

Dwight L Bolinger. 1949. The sign is not arbitrary. *Thesaurus*, 1(1):52–62.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Comput. Linguist.*, 18(1):31–40.

Andy Clark. 1998. *Magic words: how language augments human computation*, pages 162–183. Cambridge University Press.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8):2149–2169.

Johannes Dellert and Gerhard Jäger. 2017. NorthEuraLex (version 0.9). *Eberhard-Karls University Tübingen: Tübingen*.

Mark Dingemanse. 2012. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10):654–672.

Mark Dingemanse. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa: A Journal of General Linguistics*, 3(1).

Mark Dingemanse, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603 – 615.

Umberto Eco. 1995. *The Search for the Perfect Language (The Making of Europe)*. Wiley-Blackwell.

John Rupert Firth. 1930. Speech [reprinted in The Tongues of Men & Speech, 1964].

J.R. Firth. 1964. *The tongues of men, and Speech*. Oxford University Press.

Michael Gasser, Nitya Sethuraman, and Stephen Hockema. 2005. Iconicity in expressives: An empirical investigation. *Experimental and empirical methods. Stanford, CA: CSLI Publications*.

E. Dario Gutierrez, Roger Levy, and Benjamin Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2379–2388. Association for Computational Linguistics.

Shoko Hamano. 1998. *The Sound-Symbolic System of Japanese*. ERIC.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

C. 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.

Kent Johnson. 2004. On the systematicity of language and thought. *Journal of Philosophy*, 101(3):111–139.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294. Association for Computational Linguistics.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Nelson F. Liu, Gina-Anne Levow, and Noah A. Smith. 2018. Discovering phonesthemes with sparse regularization. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 49–54. Association for Computational Linguistics.

Gary Lupyan and Sharon L Thompson-Schill. 2012. The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of experimental psychology. General*, 141(1):170–186.

Margaret Magnus. 2000. *What's in a Word? Evidence for Phonosemantics*. Ph.D. thesis, Norwegian University of Science and Technology.

Alec Marantz and Morris Halle. 1993. Distributed morphology and the pieces of inflection. *The view from Building*, 20:1–52.

Hans Marchand. 1959. Phonetic symbolism in english wordformation. *Indogermanische Forschungen*, 64:146.

André Martinet. 1949. La double articulation linguistique. *Travaux du Cercle linguistique de Copenhague*, 5:30–37.

Keith Michael McCune. 1985. *The internal structure of Indonesian roots*. Ph.D. thesis, University of Michigan.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An analysis of neural language modeling at multiple scales. *CoRR*, abs/1803.08240.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Padraic Monaghan, Morten H Christiansen, and Nick Chater. 2007. The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, 55(4):259–305.

Padraic Monaghan, Morten H. Christiansen, and Stanka A. Fitneva. 2011. The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3):325–347.

Padraic Monaghan, Gary Lupyan, and Morten Christiansen. 2014a. The systematicity of the sign: Modeling activation of semantic attributes from nonwords. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Padraic Monaghan, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. 2014b. How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369:20130299.

Katya Otis and Eyal Sagi. 2008. Phonaesthemes: A corpus-based analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Geoffrey K Pullum and Barbara C Scholz. 2007. Systematicity and natural language syntax. *Croatian Journal of Philosophy*, 7(21):375–402.

M.S. Ratliff. 1992. *Meaningful Tone: A Study of Tonal Morphology in Compounds, Form Classes, and Expressive Phrases in White Hmong*. Monograph Series on Southeast Asia, Special Report (1992) Series. Northern Illinois University, Center for Southeast Asian Studies.

Richard Rhodes. 1981. On the semantics of the Ojibwa verbs of breaking. *Algonquian Papers-Archive*, 12.

Ferdinand de Saussure. 1912. Adjectifs indo-européens du type caecus "aveugle". In *Festschrift Vilhelm Thomsen zur Vollendung des siebzigsten Lebensjahres am 25. Januar 1912, dargebracht von Freunden und Schülern*, pages 202–206. Leipzig: Otto Harrassowitz. Reprinted in Saussure 1922: 595–599.

Ferdinand de Saussure. 1916. *Course in General Linguistics*. Columbia University Press. English edition of June 2011, based on the 1959 translation by Wade Baskin.

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*, pages 53–56.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc.

Andrew Spencer. 1991. *Morphological theory: An introduction to word structure in generative grammar*, volume 2. Basil Blackwell Oxford.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

John Wallis. 1699. Grammar of the English language.

Sren Wichmann, Eric W. Holman, and Cecil H. Brown. 2018. The ASJP database (version 18).

John Wilkins. 1668. *An essay towards a real character, and a philosophical language*. Gellibrand.

A NorthEuraLex Results

Language	H(W)	U($W \mid V$)	U($W \mid V$;POS)
abk	2.8432	**1.76%**	-0.26%
ady	3.2988	**2.00%**	0.50%
ain	3.0135	0.54%	-0.50%
ale	2.5990	**1.38%**	0.47%
arb	3.0872	**1.74%**	-0.07%
ava	2.8161	**2.55%**	-0.22%
azj	3.0713	**1.68%**	**1.42%**
bak	3.0652	**2.17%**	0.44%
bel	3.1212	**1.48%**	-0.37%
ben	3.2638	**1.69%**	**0.65%**
bre	3.1430	**0.57%**	**1.43%**
bsk	3.4114	0.17%	0.10%
bua	2.8739	**1.94%**	0.02%
bul	3.2150	**1.63%**	0.19%
cat	3.1536	**1.75%**	0.11%
ces	3.1182	**1.74%**	0.19%
che	3.2381	-1.60%	0.62%
chv	3.1185	0.43%	**0.91%**
ckt	2.8968	**1.60%**	0.47%
cym	3.2752	**1.42%**	**0.86%**
dan	3.2458	**0.66%**	0.57%
dar	3.2124	**1.93%**	-0.37%
ddo	3.2711	**2.15%**	-0.04%
deu	2.9596	**1.27%**	**0.90%**
ekk	2.9575	**0.69%**	-1.55%
ell	2.9141	0.15%	**0.89%**
enf	3.0470	**3.03%**	0.80%
eng	3.2126	**0.88%**	**0.70%**
ess	2.7369	**1.42%**	0.29%
eus	3.0070	**0.71%**	-0.57%
evn	2.8434	**1.34%**	0.64%
fin	2.8996	**1.32%**	0.23%
fra	3.3423	**1.17%**	-0.32%
gld	2.9055	**2.31%**	0.26%
gle	3.1450	0.51%	-0.36%
heb	3.1407	**1.26%**	0.79%
hin	3.0240	**1.11%**	**0.68%**
hrv	3.0776	**2.04%**	0.43%
hun	3.2520	0.44%	0.09%
hye	3.3416	**1.84%**	0.38%
isl	3.0386	0.50%	-0.71%
ita	2.8409	**2.18%**	0.57%
itl	3.4332	**1.96%**	0.27%
jpn	2.8157	**1.72%**	0.53%
kal	2.5255	**1.34%**	0.02%
kan	2.8412	0.23%	0.40%
kat	3.1831	**2.04%**	**1.06%**
kaz	3.0815	**2.19%**	-0.13%
kca	2.8779	**2.93%**	**1.40%**
ket	3.3202	**0.72%**	0.30%
khk	2.9746	0.57%	0.45%
kmr	3.1292	**2.22%**	0.26%
koi	3.2419	0.57%	0.25%
kor	3.1600	**1.66%**	0.40%
kpv	3.1685	**1.71%**	0.48%
krl	2.8655	**2.19%**	-0.71%
lat	2.8102	**1.36%**	0.01%
lav	2.8679	0.60%	-0.10%
lbe	3.0239	**0.94%**	-0.41%
lez	3.3717	**3.34%**	0.24%
lit	2.8086	**1.45%**	-1.33%
liv	3.0825	**1.11%**	-1.34%
mal	2.6773	**1.90%**	0.38%
mdf	2.9186	**1.24%**	-0.07%
mhr	2.9952	**1.08%**	**1.20%**
mnc	2.5750	**3.05%**	-0.03%
mns	2.8001	**1.03%**	0.18%
mrj	3.1771	**1.74%**	0.49%
myv	2.8785	**1.61%**	**0.75%**
nio	2.8985	**1.96%**	**1.46%**
niv	3.4408	**1.46%**	0.45%
nld	3.0407	**1.56%**	-0.40%
nor	3.0315	**0.68%**	0.21%
olo	3.0151	**1.38%**	0.49%
oss	3.2484	**1.42%**	-0.45%
pbu	3.2840	**1.58%**	-0.05%
pes	2.8443	**1.63%**	-0.17%
pol	3.3167	**1.65%**	0.27%
por	3.2509	**1.19%**	0.10%
ron	3.3667	0.43%	-0.99%
rus	3.3538	**1.88%**	0.17%
sah	3.0002	-1.29%	-0.37%
sel	2.8460	**1.86%**	0.76%
sjd	2.7920	-0.05%	0.30%
slk	3.1928	**1.27%**	0.46%
slv	2.8685	**2.13%**	-0.40%
sma	2.5011	**2.02%**	-0.14%
sme	2.6746	**2.10%**	-0.17%
smj	2.5975	**0.86%**	-0.52%
smn	2.9281	**1.50%**	0.22%
sms	2.7608	**1.06%**	-0.56%

Language	H(W)	U($W \mid V$)	U($W \mid V$;POS)
spa	2.9777	**1.91%**	**2.07%**
sqi	3.3473	0.22%	**0.69%**
swe	2.8600	**0.64%**	-0.44%
tam	2.6851	-0.19%	-0.63%
tat	3.1365	**1.50%**	0.17%
tel	2.8458	0.06%	-1.34%
tur	2.9646	**1.93%**	0.81%
udm	3.1042	**2.72%**	0.37%
ukr	3.1135	**1.46%**	0.48%
uzn	3.0624	**1.26%**	0.13%
vep	3.2055	**2.53%**	**1.21%**
xal	3.2090	**1.50%**	0.51%
ykg	2.9680	**1.79%**	0.65%
yrk	2.8453	**1.97%**	0.49%
yux	3.0704	-0.29%	-0.18%

Table 3: NorthEuraLex languages and p-values of systematicity. Bold entries are statistically significant at $p < 0.05$, after Benjamini–Hochberg (1995) correction.

Language	H(W)	U(W;V)	U(W;$V \mid$ POS)
abk	2.8432	**0.0500**	-0.0071
ady	3.2988	**0.0661**	0.0158
ain	3.0135	0.0161	-0.0150
ale	2.5990	**0.0358**	0.0117
arb	3.0872	**0.0538**	-0.0020
ava	2.8161	**0.0717**	-0.0059
azj	3.0713	**0.0517**	**0.0429**
bak	3.0652	**0.0666**	0.0130
bel	3.1212	**0.0462**	-0.0110
ben	3.2638	**0.0553**	**0.0206**
bre	3.1430	**0.0181**	**0.0444**
bsk	3.4114	0.0057	0.0034
bua	2.8739	**0.0558**	0.0007
bul	3.2150	**0.0523**	0.0060
cat	3.1536	**0.0550**	0.0032
ces	3.1182	**0.0543**	0.0055
che	3.2381	-0.0519	0.0194
chv	3.1185	0.0135	**0.0282**
ckt	2.8968	**0.0464**	0.0131
cym	3.2752	**0.0464**	**0.0275**
dan	3.2458	**0.0214**	0.0183
dar	3.2124	**0.0621**	-0.0114
ddo	3.2711	**0.0702**	-0.0013
deu	2.9596	**0.0377**	**0.0261**
ekk	2.9575	**0.0203**	-0.0438
ell	2.9141	0.0044	**0.0252**
enf	3.0470	**0.0923**	0.0233
eng	3.2126	**0.0284**	**0.0226**
ess	2.7369	**0.0388**	0.0076
eus	3.0070	**0.0214**	-0.0166
evn	2.8434	**0.0382**	0.0175
fin	2.8996	**0.0384**	0.0063
fra	3.3423	**0.0392**	-0.0104
gld	2.9055	**0.0670**	0.0073
gle	3.1450	0.0161	-0.0111
heb	3.1407	**0.0396**	0.0243
hin	3.0240	**0.0336**	**0.0200**
hrv	3.0776	**0.0627**	0.0127
hun	3.2520	0.0143	0.0029
hye	3.3416	**0.0615**	0.0125
isl	3.0386	0.0153	-0.0208
ita	2.8409	**0.0618**	0.0153
itl	3.4332	**0.0674**	0.0090
jpn	2.8157	**0.0485**	0.0141
kal	2.5255	**0.0340**	0.0005
kan	2.8412	0.0066	0.0111
kat	3.1831	**0.0649**	**0.0325**
kaz	3.0815	**0.0676**	-0.0039
kca	2.8779	**0.0843**	**0.0387**
ket	3.3202	**0.0240**	0.0100
khk	2.9746	0.0170	0.0128
kmr	3.1292	**0.0694**	0.0078
koi	3.2419	0.0185	0.0077
kor	3.1600	**0.0524**	0.0122
kpv	3.1685	**0.0542**	0.0148
krl	2.8655	**0.0629**	-0.0195
lat	2.8102	**0.0381**	0.0002
lav	2.8679	0.0172	-0.0027
lbe	3.0239	**0.0285**	-0.0119
lez	3.3717	**0.1126**	0.0077
lit	2.8086	**0.0409**	-0.0354
liv	3.0825	**0.0342**	-0.0401
mal	2.6773	**0.0508**	0.0097
mdf	2.9186	**0.0363**	-0.0021
mhr	2.9952	**0.0325**	**0.0348**
mnc	2.5750	**0.0785**	-0.0006
mns	2.8001	**0.0289**	0.0048
mrj	3.1771	**0.0552**	0.0151
myv	2.8785	**0.0463**	**0.0208**
nio	2.8985	**0.0569**	**0.0408**
niv	3.4408	**0.0504**	0.0147
nld	3.0407	**0.0474**	-0.0118
nor	3.0315	**0.0206**	0.0061

Language	H(W)	U($W;V$)	U($W;V \mid$ POS)
olo	3.0151	**0.0415**	0.0143
oss	3.2484	**0.0460**	-0.0140
pbu	3.2840	**0.0518**	-0.0017
pes	2.8443	**0.0463**	-0.0046
pol	3.3167	**0.0547**	0.0086
por	3.2509	**0.0387**	0.0031
ron	3.3667	0.0144	-0.0322
rus	3.3538	**0.0631**	0.0056
sah	3.0002	-0.0388	-0.0111
sel	2.8460	**0.0528**	0.0207
sjd	2.7920	-0.0013	0.0082
slk	3.1928	**0.0406**	0.0139
slv	2.8685	**0.0611**	-0.0111
sma	2.5011	**0.0505**	-0.0033
sme	2.6746	**0.0562**	-0.0043
smj	2.5975	**0.0223**	-0.0129
smn	2.9281	**0.0439**	0.0061
sms	2.7608	**0.0292**	-0.0149
spa	2.9777	**0.0568**	**0.0599**
sqi	3.3473	0.0073	**0.0226**
swe	2.8600	**0.0182**	-0.0124
tam	2.6851	-0.0050	-0.0167
tat	3.1365	**0.0471**	0.0050
tel	2.8458	0.0017	-0.0374
tur	2.9646	**0.0574**	0.0234
udm	3.1042	**0.0843**	0.0110
ukr	3.1135	**0.0456**	0.0142
uzn	3.0624	**0.0386**	0.0039
vep	3.2055	**0.0812**	**0.0374**
xal	3.2090	**0.0482**	0.0156
ykg	2.9680	**0.0532**	0.0186
yrk	2.8453	**0.0561**	0.0133
yux	3.0704	-0.0088	-0.0054

Table 4: NorthEuraLex languages and their uncertainty coefficients. Bold entries are statistically significant at $p < 0.05$, after Benjamini–Hochberg (1995) correction.