COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2016) 1–21

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12453

Wordform Similarity Increases With Semantic Similarity: An Analysis of 100 Languages

Isabelle Dautriche,[a,b] Kyle Mahowald,[c] Edward Gibson,[c] Steven T. Piantadosi[d]

[a]*Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University*
[b]*School of Philosophy, Psychology and Language Sciences, The University of Edinburgh*
[c]*Department of Brain and Cognitive Science, MIT*
[d]*Department of Brain and Cognitive Sciences, University of Rochester*

Received 18 May 2015; received in revised form 15 September 2016; accepted 19 September 2016

Abstract

Although the mapping between form and meaning is often regarded as arbitrary, there are in fact well-known constraints on words which are the result of functional pressures associated with language use and its acquisition. In particular, languages have been shown to encode meaning distinctions in their sound properties, which may be important for language learning. Here, we investigate the relationship between semantic distance and phonological distance in the large-scale structure of the lexicon. We show evidence in 100 languages from a diverse array of language families that more semantically similar word pairs are also more phonologically similar. This suggests that there is an important statistical trend for lexicons to have semantically similar words be phonologically similar as well, possibly for functional reasons associated with language learning.

Keywords: Lexicon; Phonology; Semantics; Lexical design; Arbitrariness of the sign

1. Introduction

Why do languages have the set of words that they do? Although the set of words chosen by any language is often regarded as arbitrary (Hockett, 1960; de Saussure, 1916), there are in fact well-established regularities in lexical systems which are the result of functional pressures associated with language use and its acquisition. Most notably, across languages, words that need to be communicated about more frequently tend to be short

Correspondence should be sent to Isabelle Dautriche, School of Philosophy, Psychology and Language Sciences, University of Edinburgh. E-mail: isabelle.dautriche@gmail.com

(Zipf, 1949), predictable (Piantadosi, Tily, & Gibson, 2011) and simple (Dautriche, Mahowald, Gibson, Christophe & Piantadosi, 2014). Patterns can also be found in which specific wordforms are in a language, including the presence of clusters of phonological forms (over and above effects of phonotactics or morphology) that may convey an advantage for learning, memory and lexical retrieval (Dautriche et al., 2014).

Besides properties of wordforms, another important property of lexicons is how these wordforms are associated with meanings. While a certain degree of arbitrariness in form-meaning mappings is apparent as different languages pick different wordforms to refer to the same meaning (e.g., the word for *dog* is "chien" in French, "pero" in Spanish and "kutta" in Hindi), the particular sets of form-meaning mappings chosen by any given language may also be constrained by a number of competing pressures associated with language use and its acquisition (see Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015, for review).

Several studies suggest that regularities in form-meaning mappings may facilitate individual word learning (in the case of iconic words, e.g., Imai & Kita, 2014) but also support the learning of categories and the generalization to novel words (e.g., Monaghan, Christiansen, & Fitneva, 2011). The idea is that learning similarities among referents (and hence forming semantic categories) may be facilitated if these similarities appear at the level of the wordform. For instance, it might be easier to learn the association of phonologically similar words, like *fep* and *feb*, to CAT and DOG than to CAT and UMBRELLA. This advantage in learning predicts that phonologically similar words would tend to be more semantically similar over the lexicon as a whole. In this spirit, compressible languages may be preferred by learners because they allow for compressible, thus simpler, representations that are easier to learn (Chater & Vitányi, 2003; Kemp & Regier, 2012). Previous work studying the evolution of languages in the laboratory shows that, indeed, the process of language transmission leads to lexicons that are increasingly ambiguous where a single wordform tends to be reused to express several meanings, making the language easier to learn (Kirby, Cornish, & Smith, 2008). More generally, it should be easier to learn languages whose words tend to sound similar to each other because there is less phonetic material to learn, remember, or produce (Gahl, Yao, & Johnson, 2012; Stemberger, 2004; Storkel, Armbruster, & Hogan, 2006; Storkel & Lee, 2011; Vitevitch & Sommers, 2003).

Yet there may also be functional disadvantages for form-meaning regularities that may limit the tendency for lexicons to form strong associations between phonological and semantic form. From an information theory perspective, communication can be modeled as the transfer of information over a noisy channel (Shannon, 1948). Successful communication requires listeners to be able to recover what is being said; therefore, the likelihood that two distinct words would be confusable should be minimized according to this constraint. There is much evidence that perceptual distinctiveness plays an important role in shaping the phonology of languages (e.g., Flemming, 2004; Graff, 2012; Lindblom, 1986; Wedel, Kaplan, & Jackson, 2013) and in driving speakers' production (Aylett & Turk, 2004; Bell et al., 2003; Levy & Jaeger, 2007). However, one important feature of semantically related words is that they are likely to occur in similar contexts. For

instance, weather words like "rain," "wind," and "sun" are all likely to occur in the same discourse contexts—namely, when people are talking about the weather. As a result, one might imagine that context makes it more difficult to distinguish between semantically and phonologically similar words. If someone said, "Weather forecast: today and tomorrow" the missing word could plausibly have be "sun" or "wind," but it is unlikely to be "boat" or "John." Because the context constrains the possible words in the sentence, the set of possible words for a given context should be maximally distinct in order to be recognized accurately. These considerations are present while learning the individual meanings of words (as opposed to categories), as it has been shown that toddlers find it difficult to learn words that are similar both across phonological and semantic dimensions (Dautriche, Swingley, & Christophe, 2015). Therefore, information-theoretic considerations would predict that semantically related words should be more distant in phonological space than semantically unrelated words.

In this work, we investigate whether the relationship between semantic distance and phonological distance is non-arbitrary and in which direction. If there is a positive correlation between semantic distance and phonological distance—that is, more similar wordforms are more semantically similar—above what would be expected by chance alone, then this would imply a pressure for phonological clustering that is tied specifically to meaning. On the other hand, if there is a negative correlation between semantic distance and phonological distance below chance level, there would be a pressure for words' meanings to be more distinct relative to phonological distance, likely due to communicative pressures of confusability.

Certainly, there are well-known exceptions to arbitrariness: sound symbolism, in which languages encode iconic correspondences between sounds and meanings,[1] is one such form-meaning regularity and is present across many languages and cultures (e.g., Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Bremner et al., 2013; Childs, 1994; Hamano, 1998; Kim, 1977). For instance, contrasts in vowels have been shown to correspond to contrasts in magnitude (e.g., /i/ vs. /a/; "this" vs. "that" in English) and repetition of syllables to words describing iterative events (e.g., "barbar" *all the time* in Hindi). It has been shown that some of these encodings can be retrieved without prior linguistic experience: For instance, in a meaning-guessing task, adults guess that "buba" refers to round shapes and "kiki" refers to spiky shapes (the "buba-kiki" effect, e.g., Bremner et al., 2013). Relatedly, certain sequences of sounds, called phonesthemes, tend to carry a certain semantic connotation. For instance, there is a tendency in English for *gl*-words to be associated with light reflectance as in "glitter," "glimmer," and "glisten" (Bergen, 2004; Bloomfield, 1933) or words ending with *-ack* and *-ash* associated with abrupt contact (e.g., "smack," "smash," "crash," "mash"). Additionally, certain meaning distinctions may be present in the phonological form of words more systematically. For instance, semantic features, such as objects versus actions, that are associated with grammatical distinctions may be marked phonologically (Monaghan & Christiansen, 2008; Pinker, 1984). Yet it is unclear whether these constitute exceptions or are representative of a more systematic pattern.

Previous studies (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Shillcock, Kirby, McDonald, & Brew, 2001; Tamariz, 2008) have examined the correlation between semantic distance and phonological distance in English and Spanish. In these works, the authors found that phonologically similar words tend to be more semantically similar, even in a small subset of the lexicon consisting of monomorphemic words in which etymological entries have been removed, suggesting that such correlation, exist over and beyond morphology and historical variants. Importantly, Monaghan et al. (2014) showed that, in English, such a systematic pattern may not be the result of localized pockets of form-meaning systematicity (e.g., sound symbolism, phonesthemes) but is diffusely distributed across the whole lexicon (see, however, the recent study of Gutiérrez, Levy, and Bergen, 2016, for the finding that much of this systematicity is concentrated in such localized form-meaning patterns when using a different method). While these results are telling, the sample of one or two languages does not indicate if form-meaning regularities in the lexicon are the result of a learnability pressure that universally applies or are historical accidents of the languages studied.

The existence of large-scale datasets in a large number of languages now makes it possible to investigate semantic and phonological relatedness across human languages more generally. In the present work, we use a dataset of 100 languages extracted from Wikipedia from a diverse array of language families. First, we performed several statistical tests to look at the correlation between semantic similarity (calculated using latent semantic analysis [LSA] over each Wikipedia corpus) and orthographic similarity (a proxy for wordform similarity, since phonological codes were not available for most languages in the dataset). Second, because the presence of a correlation may be explained by a small number of words rather than a systematic pattern in the vocabulary, we further probed the relation between semantic and phonological similarity, using a different measure looking at the relationship between semantic relatedness and the likelihood of finding a minimal pair. Because our results are limited by the absence of phonemic representations and the inclusion of morphologically complex words, we also used a subset of four languages to assess whether the correlation between semantic and phonological similarity still holds in a set of monomorphemic words with phonemic representations. All our analyses suggest that semantically similar words tend to be phonologically similar, thus providing large-scale, crosslinguistic evidence for phonological clustering of semantically similar words.

2. Methods

2.1. 100 orthographic lexicons from Wikipedia

We extracted the lexicons of 100 languages from the Wikipedia database. These languages span a total of 49 language families, making it possible to investigate how semantic and wordform similarities relate cross-linguistically. We define as the lexicon of these languages the 5,000 most frequent wordforms in the Wikipedia corpus.[2] Because a proper lemmatizer does not exist for most of these languages, all of the 5,000 most frequent

wordforms were included regardless of their morphemic status. In order to minimize the impact of semantic similarity due to morphological regularity (e.g., while comparing "cat" and "cats" or "big" and "biggest") and to avoid comparing long words (which are unlikely to be similar by chance alone) to short words (which are necessarily closer together in phonetic space), we only compared words of the same length. Because no phonemic transcription is available in Wikipedia, our measure of wordform similarity relied on the orthographic scripts of these languages. Though this remains an approximation, the number of characters in a word is reliably correlated with the number of phonemes in Dutch ($r = .87$), English ($r = .83$), German ($r = .89$), and French ($r = .79$; all $p < .001$), which are the languages for which we have phonemic and orthographic codes available (see below).

2.2. Four phonemic lexicons

To assess whether a correlation between semantic similarity and phonological similarity holds in a set of monomorphemic words with phonemic representations, we also used phonemic lexicons derived from CELEX for Dutch, English, and German (Baayen, Piepenbrock, & Gulikers, 1995) and Lexique for French (New, Pallier, Brysbaert, & Ferrand, 2004). The lexicons were restricted to include only monomorphemic lemmas (coded as "M" in CELEX; I.D. (a French native speaker) identified mono-morphemes by hand for French. That is, monomorphemes were identified as containing neither inflectional affixes (like plural *-s*) nor derivational affixes like *-ness*. Diphthongs were transformed into two-character strings in order to capture similarity among their component vowels. In each lexicon, we removed a small set of words containing foreign characters and removed stress marks as well as compound words. In order to focus on the most used parts of the lexicon, we selected only words whose frequency in CELEX or in Lexique is strictly >0.[3] Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., "bat"), we included this form only once. This criterion excluded spelling variants of the same word (e.g., "analyze"/"analyse") that would have been considered as minimal pairs when using orthographic codes. This resulted in a lexicon of 5,459 words for Dutch, 6,512 words for English, 4,219 words for German, and 6,782 words for French. The resulting lexicons are available at https://osf.io/rvg8d/)

2.3. Variables under consideration

For each pair of words of the same length in each of the lexicons, we computed the pair's:

- *Orthographic/Phonological distance:* Because we compared words of the same length to reduce the effect of morphology in the 100 orthographic lexicons, we used the raw edit distance, or raw Levenshtein distance between two orthographic strings in the case of the 100 orthographic lexicons and phonemic strings in the case of the four phonemic lexicons. The smaller the distance, the more similar

wordforms are to each other. For example, the words "cat" and "car" are very similar, with an edit distance of 1. Note that for all languages, a character is taken as a single unit (even for abjads, e.g., Arabic, or abugidas, e.g., Tai). Furthermore, all languages in the sample use alphabetic scripts (see Appendix A).

- *Semantic distance:* We used LSA (Landauer & Dumais, 1997), a class of distributional semantic models that build on the hypothesis that words' meanings can be inferred from their context (Harris, 1954) where context is assumed to be a "bag of words."[4] Under this approach, two words are expected to be semantically similar if their pattern of co-occurence in some observed text is similar. For example, "cat" and "dog" will be more similar than "cat" and "bottle" because they are more likely to co-occur in the same texts. One advantage of using this technique as a proxy for semantics rather than hand-made lexical taxonomies such as WordNet (Miller, 1995)—which is only extensively developed in English—is that it can be adapted for any language given a sufficiently large corpus. We note, however, that the results obtained from several measures of word distance using WordNet provide the same results as an LSA model trained on English (see Appendix B).

We applied LSA to Wikipedia for each language using the Gensim package (Rehurek & Sojka, 2010) in Python. This model splits the whole Wikipedia corpus into documents consisting of n lines of text and constructs a word-document matrix where each row i is a word and each column j, a document. Each matrix cell c_{ij} corresponds to the frequency count of word i in document j. The matrix is then reduced to a dimension d corresponding to the number of semantic dimensions of the model using Singular Value Decomposition. The semantic distance between two words is computed as 1 minus the absolute value of the cosine of the angle between the two word vectors in the space of dimension d. A value close to 0 indicates that two words are close in meaning, whereas values close to 1 indicate that the meanings are not related.

For our purposes we defined a document as a Wikipedia article (number of documents per language corpus: median = 42,989; min = 104—Buginese; max = 36.6 billion—English) and d = 500 dimensions[5] based on Fourtassi and Dupoux (2013) and Rehurek and Sojka (2010). We also discarded words that appear in fewer than 20 documents and in more than 50% of the documents to account for the fact that very common and very rare terms are weak predictors of semantic content (a procedure commonly used in Machine Learning; Luhn, 1958).

3. Results

3.1. Large-scale effects of semantics on 100 languages

3.1.1. Pearson correlations analysis

To analyze the relationship between semantic distance and orthographic distance, we computed Pearson correlations between the semantic distance of all pairs of words of the

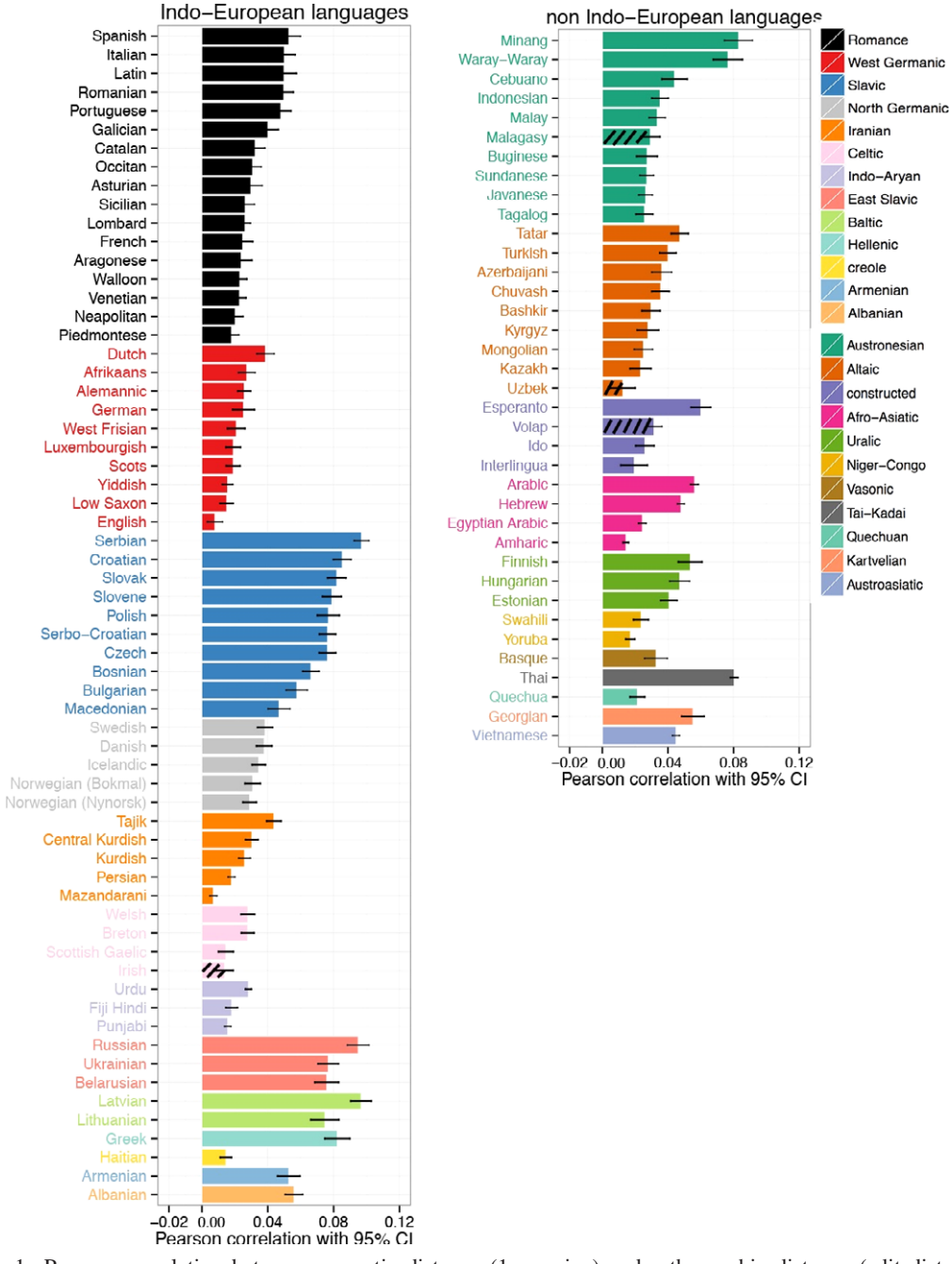

Fig. 1. Pearson correlation between semantic distance (1 - cosine) and orthographic distance (edit distance) for each language for word of length 4. Languages are grouped per language family for Indo-European languages (left plot) and non Indo-European languages (right plot). Solid bars are used for significant Pearson correlations ($p < .001$) and striped bars for non-significant correlations ($p > .001$) compared to the chance level estimated from 10,000 meaning rearrangements.

Table 1
For each length: (a) the mean Pearson correlation across languages for the relationship between semantic and orthographic distance; (b) the proportion of languages that show a positive correlation between semantic distance and orthographic distance; and (c) the proportion of languages for which this relationship is significantly different from chance at $p < .001$, chance being the correlation obtained during 10,000 random form-meaning reassignments

Word Length	Mean Correlation	Proportion Showing Positive Correlation	Proportion Showing Significant Correlation
3 letters	0.049	0.99	0.84
4 letters	0.04	1	0.94
5 letters	0.037	1	0.99
6 letters	0.039	1	0.99
7 letters	0.045	1	0.99

same length (focusing on words of length 3–7) and the pairs' orthographic distance. To evaluate the correlation between semantic distance and orthographic distance, we need to compare it to a baseline that reflects the chance correlation between form and meanings in the lexicon. We created such a baseline by randomly permuting the form-meaning mappings for words of a given length, randomly reassigning every word meaning to a word of the same length. For example, the meaning of "car" could be reassigned to "cat" and the meaning of "dog" to "rat." Under this permutation, the mapping between form and meaning (unlike in the real lexicon) is entirely arbitrary for words of a given length. For each language, we randomly reassigned meanings 10,000 times and computed Pearson correlations for each word length. Because the sampling distribution of the Pearson correlation is not normally distributed, we transformed each correlation to fit this assumption (Fisher *Z*-transformation: $z = 0.5 \times (\ln(1 + r) - \ln(1 - r))$, where r stands for the Pearson correlation). We then asked whether the transformed correlation between word-form distance and semantic distance of the real lexicons falls outside the range of transformed correlation values that could be expected by chance, where chance means random form-meaning assignments.

Fig. 1 summarizes this hypothesis test for four-letter words across the 100 languages. Each bar represents the Pearson correlation score for a given language, and each color represents a language family. We observe that (a) all correlations are positive; (b) most of the correlations are significantly positive (in 94/100 languages; solid bars), meaning that as semantic distance increases, orthographic distance increases as well.

We computed a z-score using the mean and standard deviation of the transformed correlation scores estimated from the 10,000 meaning rearrangements. The p-value reflects the probability that the real lexicon correlations could have arisen by chance. As can be seen in Table 1, we found that the great majority of languages display a significant positive correlation between semantic distance and orthographic distance for all lengths.[6] The observation that the correlation is significant for long words in almost all languages is somewhat unsurprising given that many of these words may be morphologically complex. But importantly, this correlation appears across most languages even for words of small

Table 2
Summary of the full models including random intercepts and slopes for language, subfamily, and family for edit distance for each word length.

Word Length	Edit Distance
3 letters	0.10*
4 letters	0.06*
5 letters	0.04*
6 letters	0.04*
7 letters	0.04*

*By a likelihood test, the predictor significantly improves model fit at $p < .001$.

lengths (3–4) which are less likely to be morphologically complex. Yet, even though the correlations are generally highly significant, one needs to observe that these are tiny effects, explaining only a very small amount of the variance ($r < .05$)—a point we return to in the General discussion.

3.1.2. Mixed effect analysis

To ensure that the observed effect holds when controlling for the variation among language families, we ran a mixed effect regression predicting scaled semantic distance for each pair of words from the edit distance between the words of the pair. We used a maximal random effect structure with random intercepts for each language, language subfamily, and language family and slopes for edit distance for each of those random intercepts. Because of the large number of data points, we fit each length separately (words of length 3 through length 7). The semantic distance was centered around the mean semantic distance for each length and each language and scaled by the standard deviation for each length and each language. We compared the full model to an identical model without a fixed effect for the number of minimal pairs using a likelihood ratio test.

Table 2 shows the coefficient estimates for an effect of edit distance on semantic distance. For every word length, the coefficient for edit distance is significantly positive, meaning that increased semantic distance comes with increased edit distance beyond effects of language family or subfamily.

3.2. Likelihood of finding a minimal pair in 100 languages

The observation of a correlation between semantic similarity and orthographic similarity may be driven by a consistent effect in the lexicon (e.g., a change in orthographic distance yields a change in semantic distance) but may be the by-product of a small fraction of words that drive the correlation (e.g., pockets of sound symbolic words). To get around this issue, we looked at the relationship between semantic relatedness and the likelihood of finding a minimal pair. For each language, we compared the number of minimal pairs in the top 10% of semantically related words pairs *ntop*, and in the bottom 10% of semantically related words pairs *nbottom* by looking at the ratio $\frac{n_{top}}{n_{bottom}}$. A ratio below 1

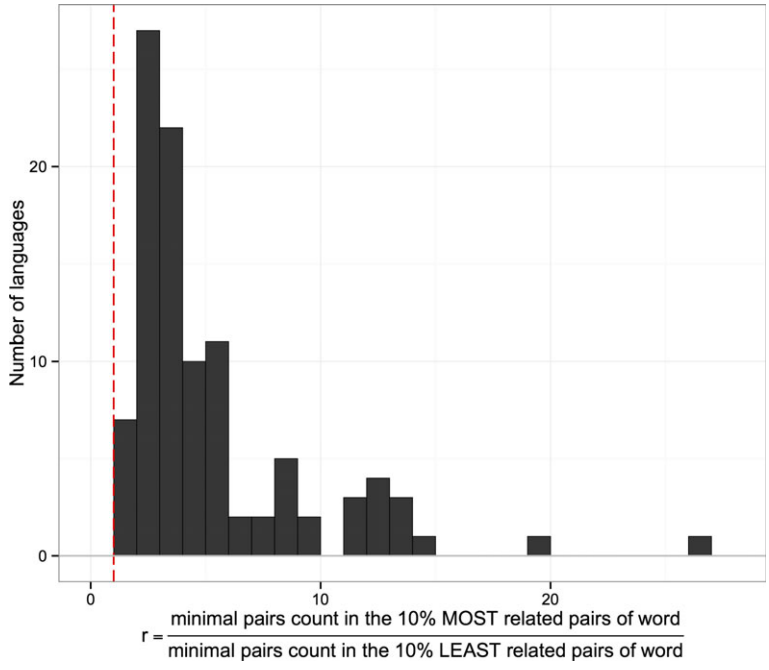

Fig. 2. Distribution of the ratio of the number of minimal pairs in the 10% most related words compared to the number of minimal pairs in the 10% least related words in a given lexicon, across all the languages. A ratio below 1 means that there are more minimal pairs in semantically unrelated words than in related words, whereas a ratio greater than 1 means that there are more minimal pairs among semantically related words than unrelated words.

mean that there are more minimal pairs in semantically unrelated words. A ratio below 1 means that there are more minimal pairs in semantically unrelated words than in related words, while a ratio greater than 1 means that there are more minimal pairs among semantically related words than unrelated words. Fig. 2 shows the histogram of the distribution of ratio $\frac{n_{top}}{n_{bottom}}$ across all languages. As we can observe, in all 100 languages, minimal pairs are on average 3.52 (median of the distribution) times more likely to appear in the top 10% semantically related words than in the least 10% related words. Note that we obtain qualitatively the same results by looking at the 25% most related and the 25% least related words or other percentages. Importantly the effect was present for words of all lengths.

3.3. Generalizing form-meaning regularity to monomorphemic words

Phonologically similar words tend also to be semantically similar across a large range of typologically different languages. We have interpreted this observation in terms of a pervasive systematic relationship between the forms and the meaning of

words; one potential alternative explanation for the positive correlation between semantic and orthographic distances is the presence of morphological regularity among the 100 lexicons we studied here. To limit this effect, we studied words of the same length and broke up our results by word length: Importantly even in short words composed of 3–4 letters, thus less likely to be morphologically complex, we found a positive correlation between phonological and semantic similarities. In addition, we conducted another analysis looking at minimal pairs and found that there are more minimal pairs among semantically related than semantically unrelated words. Yet, certainly, this does not exclude the possibility that some morphological regularity is remaining. Note also that while the Wikipedia corpus allows us to study many typologically different languages, in some languages there is no consensus for the spelling of words (e.g., Vietnamese) and Wikipedia does not adopt any spelling convention for multidialectal languages (e.g., English) such that some words may be mistakenly taken as phonologically close while being semantically identical (e.g., "analyze"/"analyse") and contribute to the correlation we observe.

To further separate the correlation between phonological and semantic distance due to morphemic regularity from the correlation we are interested in, we restricted our analysis to four languages, Dutch, English, French, and German, for which monomorphemic codes are readily available. Because we used the phonemic codes of these lexicons, spurious effects due to a lack of spelling convention in these languages could be eliminated.

For the monomorphemes of these four languages, we computed the transformed Pearson correlations between semantic distance and phonological distance for each word length and compared it to the correlations obtained after 10,000 random form-meaning reassignments. As shown in Fig. 3, the correlations obtained in the real lexicons for each word length (the red dot) tend to be significantly more positive than the correlations obtained in 10,000 random configurations of form-meaning pairings (the histograms).

Overall semantic distance is positively correlated with phonological distance ($r = .04$) significantly more than what would be expected by chance ($p < .001$ across most lengths and all languages). Careful observation of the monomorphemes extracted from CELEX revealed that some polymorphemic words slipped into the monomorphem list. To eliminate potential issues with the morphological coding of CELEX, we conducted the same analysis on a restricted set of polymorphemic words, monosyllabic words, for which we only took into account words composed of 3–4 phonemes as there are very few longer monosyllabic words (1,762 words for Dutch, 2,578 for English, 1,249 for French and 817 for German). While the average correlation found among monosyllabic words is smaller ($r = .02$), this is still significantly more than what would be expected by chance ($p < .001$ across most lengths and all languages). Thus, it seems unlikely that morphology similarity is causing the relationship that we see between semantic similarity and phonological similarity.

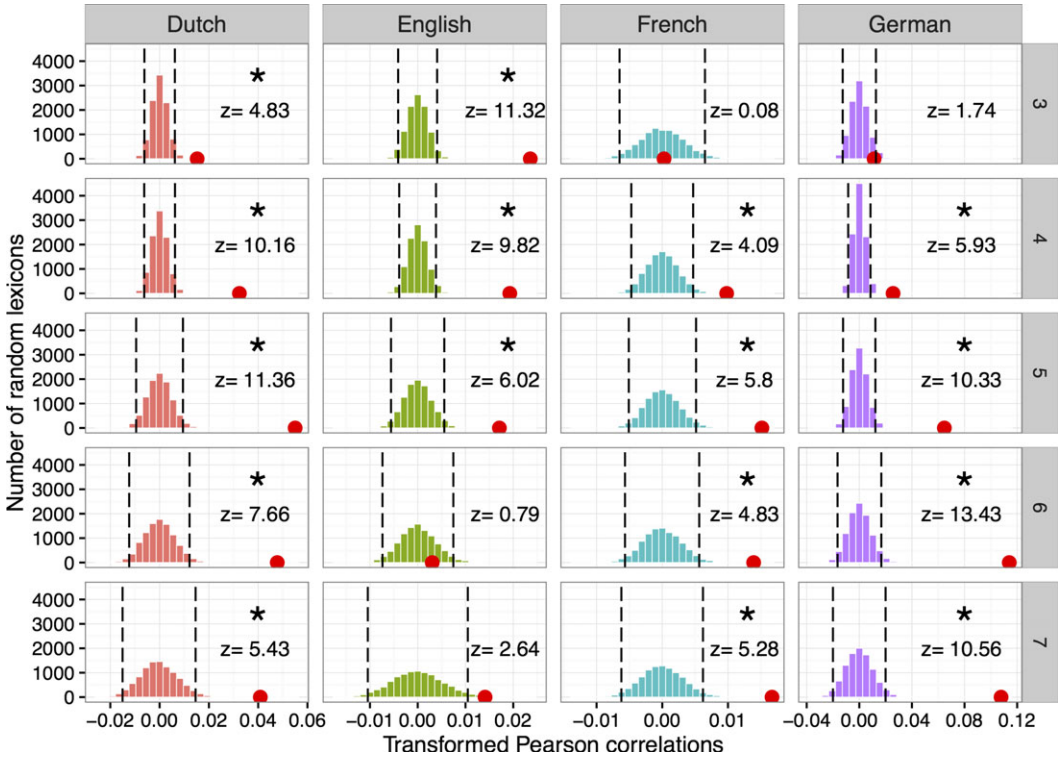

Fig. 3. Transformed Pearson correlations between semantic distance and phonological distance for words of length 3–7 (in rows) for Dutch, English, French, and German. Each histogram shows a distribution of transformed correlations obtained after 10,000 random form-meaning assignments (chance level). The red dots are the transformed correlations found in the real lexicon for that particular length. The dotted lines represent the 95% confidence interval. The *z*-statistics compare the transformed Pearson correlations in the real lexicon with the chance distribution corresponding to the distribution of transformed Pearson correlations obtained in the 10,000 random form-meaning mappings for each word length and each language. The star indicates that the correlation score is significant at $p < .001$.

4. General discussion

We have shown that across 100 languages, similar wordforms tend also to be more semantically similar above and beyond what could be expected by chance extending previous research that examined single languages (Monaghan et al., 2014; Shillcock et al., 2001; Tamariz, 2008) to a set of typologically different languages.

What could be the reasons for form-meaning regularity in the lexicon? One obvious possibility is the presence of morphologically complex words in these lexicons. In order to estimate the contribution of morphology from this correlation, we conducted several analyses. First, we analyzed our results by word length and showed that the correlation holds for short words which are less likely to be morphologically complex. Second, we

analyzed the relationship between semantic relatedness and the likelihood of finding a minimal pair and found that minimal pairs are more likely among semantically similar words than semantically distant words. Finally, we conducted a correlation analysis on the set of monomorphemic lemmas of a restricted number of languages and found a similar pattern of results. While this suggests that form-meaning regularities exist over and above morphological regularity, across a large range of typologically different languages, it is difficult to determine the extent to which morphology may be driving the effects across the 96 languages for which we could not extract monomorphemic words. Indeed, other morphological systems might be affecting performance to a greater extent than in the closely related set of Germanic and Latin languages for which we could perform the analyses on monomorphemes. For instance, in Hebrew and Arabic, morphologically related words are normally written to include their common consonantal roots, omitting the vowels. For these languages, the edit distance calculated on the orthographic transcription of words may thus be underestimated, resulting in an inflated correlation between phonological and semantic proximity. Another issue is that of languages for which the orthography contains traces of morphology which are not audible phonologically. For instance, French past tense differs in orthography for second- and third-person singular (e.g., mange*ais*/mange*ait* for the verb "eat") despite having the same pronunciation. Those phenomena may also inflate the correlation between form and meaning.[7] While these problems are related to morphophonology, morphological decomposition more generally may also be contributing substantially to the effects. Words that share morphological elements are usually also related in form and meaning (e.g., brave/bravely), yet not systematically (ginger/gingerly), indicating that the morphology-meaning relation is not perfectly reliable in languages, and this may vary for different language systems (though to our knowledge, there is no cross-linguistic study to date that investigated this relation). Therefore, although we provided here several different analyses that all point toward a positive correlation between forms and meanings above what could be expected by chance, it remains an open question the extent to which morphology, and its interaction with orthography, contribute to the observed correlations from the Wikipedia corpora.

Another possibility is that form-meaning regularity is due to etymology. Etymology is an important source of regularity in form-meaning mappings: Certain words are historically related or derived from other words in the lexicon (even when the lexicon is restricted to morphologically simple words). For example, "skirt" and "shirt" are historically the Old Norse and Old English form of the same word, whose meanings have since diverged. Previous work showed that etymological roots were not sufficient to account for the pattern of systematicity observed across the English lexicon (Monaghan et al., 2014). Future work will need to confirm whether such a global pattern of form-meaning systematicity exists over and above etymological roots for the languages under study here.

Because the correlation between forms and meanings is present across 100 languages, this suggests that systematicity of form-meaning mappings is an important statistical trend in natural languages which is the result of functional pressures, possibly associated with

learnability, that universally applies to all languages. This goes hand in hand with the recent results of Blasi et al. (2016), who showed consisted sound-symbolim in a list of 100 common words from almost two-thirds of the world's languages, supporting the idea that such a correlation may be cognitively grounded. In the early stages of word learning, having a non-arbitrary mapping between forms and meaning may be helpful to establish the mapping between the phonological and the semantic level. Indeed, Monaghan et al. (2014) showed that the first words acquired by children display more sound symbolism than later acquired words, suggesting that there may be a natural propensity to establish a cross-modal relation between form and meaning (e.g., Spector & Maurer, 2009). More systematic relations between the forms and the meaning of words may also be carried by the grammatical category of the words. Even though we looked at monomorphemes, words from the same grammatical category share phonological features (Cassidy & Kelly, 1991; Kelly, 1992), such that nouns sound more similar to other nouns and verbs to other verbs (see also Dautriche et al., 2014), and are overall more semantically closer to words of the same grammatical category (e.g., verbs are more likely to map onto actions and nouns onto objects). Such systematic form-meaning mappings may be helpful during language learning to cue grammatical categories and support generalization to novel words (Fitneva, Christiansen, & Monaghan, 2009; Monaghan et al., 2011). In addition, such regularities in the lexicon may minimize memory load by reusing similar codes and mapping them onto similar meanings, yielding lexical systems which are simpler thus easier to learn and more likely to survive the language transmission process (Brighton, Kirby, & Smith, 2005).

The present results are not consistent with the predictions of communication-based influences on lexical structure. For instance, imagine a language that displays an extreme form-meaning regularity where similar and frequent concepts such as CAT and DOG will be associated with similar wordforms such as "feb" and "fep," respectively. These words will be easily confused since their forms differ only from one phoneme and their meanings are similar. Thus, if lexicons are efficient solutions to the communicative problem of transmitting information, we expect a pressure that should disambiguate meaning and avoid words that are similar both at the semantic and the phonological level to make sure that we never misunderstand one another.

One possibility for why communicative efficiency does not seem to drive form-meaning mappings in the lexicon is that noise in the speech signal may not exert a selection pressure on form-meaning mappings in languages. Indeed, context is typically sufficient to disambiguate between meanings, since adult speakers use many cues when processing spoken sentences (e.g., prior linguistic context, Altmann & Kamide, 1999; visual information, Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; speaker, Creel, Aslin, & Tanenhaus, 2008) that have been shown to also be useful during language learning (e.g., Creel, 2012; Dautriche et al., 2015). As a result, finer-grained contextual information may be sufficient most of the time for language understanding and learning to distinguish between phonologically similar words.[8] Yet, if noise in the speech signal does not matter for efficient information transfer, it is unclear how one can explain (inter alia) why speakers modulate their production to transmit a uniform amount of information per unit of

time (see *uniform information density* effect, UID, Frank & Jaeger, 2008; Levy & Jaeger, 2007) consistent with information theory models of communication.

Another possibility is that the effect of communicative efficiency may be masked by other (functional) constraints on the lexicon. Indeed, the fact that we observe such a small correlation ($r < .05$) between the phonological forms of words and their meanings reflects the presence of a number of other constraints that influence words' phonology and semantics, possibly for communicative reasons. While the sign of the correlation suggests a greater advantage for systematic form-meaning mappings, further studies may provide evidence that communicative pressures are also at play. Finer-grained analyses suggest indeed that different phonological features may have different purposes: Tamariz (2008) shows that, in Spanish, the consonant structure of words tends to reflect systematicity of form-meaning mappings, while the vowel structure, on the contrary, shows the reverse correlation. This suggests that it is possible to achieve a trade-off between systematicity of the vocabulary and accurate word recognition (see also Zuidema [2003] for a demonstration that confusability may still be minimized while achieving a high degree of systematicity among words).

To our knowledge, with 100 languages in the sample, this is the largest cross-linguistic analysis showing a correlation between semantic similarity and phonological similarity across the whole vocabulary, suggesting the presence of a global pattern of systematicity in form-meaning mappings in languages. Because the correlation between phonological and semantic similarity is found across a wide array of typologically different languages, we propose that this pattern is the result of a functional pressure that universally applies. Certainly, a direct causal link between any functional pressure and the pattern of systematicity found in lexicons cannot be established with the present data, neither *how* such a pattern may arise in the lexicon. Research on language evolution offers a promising direction toward answering these questions: Processes of repeated language transmission have been shown to turn an arbitrary lexicon into a systematic one (e.g., Kirby et al., 2008) such that the optimal structure of the vocabulary may be one that incorporates form-meaning regularities at the large scale of the lexicon as new words are added and old ones dropped through the transmission of the lexical system.

Acknowledgments

We thank Anne Christophe, Benoit Crabbé, Emmanuel Dupoux, and all members of Tedlab, the audience at AMLaP 2014, and the audience at CUNY 2013 for helpful comments. This project is supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health (F32HD070544) as well as a post-graduate fellowship from the Fyssen to ID and an NDSEG graduate fellowship and an NSF Doctoral Dissertation Improvement Grant in linguistics to KM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Notes

1. Note that this is not specific to spoken languages; sign languages also map meanings into visual sign (see Strickland et al., 2015).
2. Since we calculated words' semantic distance for all pairs of words in the lexicon, picking only the most frequent words was done in order to limit the number of possible calculations for each language.
3. We did not use the zero-frequency words in our analysis because words with low frequency counts may lead to poor vector (semantic) representation in the Wikipedia corpus.
4. We also used another class of models using deep neural nets (word2vec), and this led qualitatively to the same results.
5. For Buginese which was the only language having fewer documents than 500 (the number of dimensions), we took $d = 20$.
6. Qualitatively similar results were observed using a Mantel test to estimate the distribution of correlations.
7. Yet we note that in the case of French, excluding these words from the lexicon, thus focusing only on the phonological form of monomorphemes, still results in a positive correlation between forms and meanings.
8. Here we modeled meanings, using context-vectors and still found a correlation. Yet it is important to note that the context we used in this case is limited to co-occurence counts with other words (a "bag of words" context) and does not include finer-grained linguistic context (e.g., syntax) or context in other modalities (e.g., visual).

References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (release 2)[cd-rom]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*, 1001.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311. doi:10.1353/lan.2004.0056

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113* (39), 10818–10823. doi:10.1073/pnas.1605782113

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

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

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution* (pp. 291–309). Oxford, UK: Oxford University Press.

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*(3), 348–369. doi:10.1016/0749-596X(91)90041-H

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Childs, G. T. (1994). African ideophones. In L. Hinton et al. (Ed), *Sound Symbolism*, (pp. 178–209). Cambridge: Cambridge University Press.

Creel, S. C. (2012). Preschoolers' use of talker information in on-line comprehension. *Child Development*, *83* (6), 2042–2056.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*(2), 633–664.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2014). Lexical clustering in efficient language design. Oral presentation in AMLaP. Edinburgh, Scotland.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, *36*(5), 967–997.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In Bruce Hayes, Robert Kirchner, and Donca Steriade (eds), *Phonetically-based phonology*, (pp. 232–276) Cambridge: Cambridge University Press.

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. *ACL*, *2013*, 165.

Frank, A., & Jaeger, T. (2008). *Speaking rationally: Uniform information density as an optimal strategy for language production. Cogsci*. Washington, DC: CogSci.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. doi:10.1016/j.jml.2011.11.006

Goslin, J., Galluzzi, C., & Romani, C. (2014). Phonitalia: A phonological lexicon for italian. *Behavior Research Methods*, *46*(3), 872–886.

Graff, P. N. H. M. (2012). Communicative efficiency in the lexicon. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Gutiérrez, E. D., Levy, R., & Bergen, B. K. (2016). Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. Proceedings of the Association for Computational Linguistics.

Hamano, S. (1998). *The sound-symbolic system of Japanese*. Gainesville: Univ. of Florida dissertation.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364. doi:10.1037/0033-295X.99.2.349

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kim, K.-O. (1977). Sound symbolism in korean. *Journal of Linguistics*, *13*(01), 67–75.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Landauer, T., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In T. Hoffman, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (pp. 849–856). Cambridge, MA: MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the International Conference on Machine Learning* (vol. 98, pp. 296–304).

Lindblom, B. (1986). Phonetic universals in vowel systems. In John J. Ohala and J. J. Jaeger (eds), *Experimental phonology*, (pp. 13–44) Orlando, FL: Academic Press.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165.

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: Subtitle-based word frequency estimates for polish. *Behavior Research Methods*, *47*(2), 471–483.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 139–164). Amsterdam: John Benjamins.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347. doi:10.1037/a0022924

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*, *369*(1651), 20130299.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

de Saussure, F. (1916). *Course in general linguistics*. New York, NY: McGraw-Hil.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656.

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

Spector, F., & Maurer, D. (2009). Synesthesia: A new approach to understanding the development of perception. *Developmental Psychology*, *45*(1), 175.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1), 413–422.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. doi:10.1044/1092-4388(2006/085)

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. doi:10.1080/01690961003787609

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, *112*(19), 5968–5973.

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

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, *31*(4), 491–504.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*(2), 179–186.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.

Zuidema, W. (2003). Optimal communication in a noisy and heterogeneous environment. In *European Conference on Artificial Life* (pp. 553–563). Springer Berlin Heidelberg.

Appendix A: Dataset of 100 lexicons from Wikipedia

Wikipedia is the largest publicly available multilingual corpus (https://dumps.wikimedia.org). The downloaded file includes the dump of all Wikipedia articles in XML format. Since all articles are not available in all languages available, the semantic content may differ across languages. We used the *Gensim* library in Python (Rehurek & Sojka, 2010) to preprocess the corpus and train a latent semantic model (LSA; Landauer & Dumais, 1997). During the preprocessing step, each Wikipedia dump is cleaned from XML syntax and then scanned for all distinct word types. The word types that appear in more than 10% of the Wikipedia articles are removed, and from the rest the 10,000 most frequent types are kept and used in the Latent Semantic model.

We started with lexicons of 115 languages from their Wikipedia databases. We excluded languages for which more than 20% of words were foreign words (usually English) among the 100 most frequent words. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the 3- to 7-letter words in Chinese Wikipedia are often English. We also excluded Korean because their orthographic length cannot compare directly with alphabetic scripts. After these exclusions, 100 languages remained. (We excluded Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, and Kannada). We analyzed the data both with and without these exclusions, and the exclusions do not significantly affect the overall direction or magnitude of the results. The languages analyzed included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, there are 12 language families represented as well as a Creole and four constructed languages (Esperanto, Interlingua, Ido, Volap) that have some speakers. (The analysis is qualitatively the same after excluding constructed languages.) The languages analyzed are shown in Tables A1 and A2.

To get a sense of how clean these Wikipedia lexicons are, we randomly sampled 10 languages for which we then inspected the 100 most frequent words and an additional 100 random words to look for intrusion of English words, HTML characters, or other undesirable properties. For the top 100 words in the lexicons of the 10 sampled languages, we found at most three erroneous words. For the same languages, we also inspected a randomly selected 100 words and found that the mean number of apparently non-intrusive words was 93.5 (with a range from 85 to 100). The most common intrusion in these languages was English words.

Table A1
Indo-European languages used (language families in bold)

Albanian: Albanian; **Armenian**: Armenian; **Baltic**: Lithuanian, Latvian; **Celtic**: Breton, Irish, Scottish Gaelic, Welsh; **creole**: Haitian; **Germanic**: Afrikaans, Alemannic, German, English, Luxembourgish, Low Saxon, Dutch, Scots, West Frisian, Yiddish; **Hellenic**: Greek; **Indo-Aryan**: Fiji Hindi, Marathi, Urdu, Bosnian, Croatian, Punjabi, Serbian; **Iranian**: Central Kurdish, Persian, Kurdish, Mazandarani, Tajik; **Italic**: Latin; **North Germanic**: Danish, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmal), Swedish; **Romance**: Aragonese, Asturian, Catalan, Spanish, French, Galician, Italian, Lombard, Neapolitan, Occitan, Piedmontese, Portuguese, Romanian, Sicilian, Venetian, Walloon; **Slavic**: Belarusian, Bulgarian, Macedonian, Czech, Polish, Russian, SerboCroatian, Slovene, Slovak, Ukrainian

Table A2
Non-Indo-European languages used (language families in bold)

Afro-Asiatic: Arabic, Amharic, Egyptian Arabic, Hebrew; **Altaic**: Mongolian, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Turkish, Tatar, Uzbek; **Austronesian**: Minang, Malagasy, Indonesian, Malay, Sundanese, Cebuano, Tagalog, Waray-Waray, Buginese, Javanese; **Austroasiatic**: Vietnamese; **Kartvelian**: Georgian; **Niger-Congo**: Swahili, Yoruba; **Quechuan**: Quechua; **Tai-Kadai**: Thai; **Uralic**: Estonian, Finnish, Hungarian; **Vasonic**: Basque; **constructed**: Esperanto, Interlingua, Ido, Volap

To have an idea on how the lexicon extracted from Wikipedia compares to the lexicon we could extract from other corpora, we extracted the 5,000 most frequent words of Wikipedia (our lexicons) to the 5,000 most frequent words in another corpora of the same language and looked at the number of words overlapping between these two. We found on average 48% (with a range from 32% to 75%) of the words overlaps in seven languages (French using the frequencies from Lexique, New et al., 2004; English, Dutch, and German from CELEX, Baayen et al., 1995; Italian from phonItalia, Goslin, Galluzzi, & Romani, 2014; Polish from Subtlex-pl, Mandera, Keuleers, Wodniecka, & Brysbaert, 2015; Spanish from Espal, http://www.bcbl.eu/databases/espal/).

Appendix B Comparison Between LSA and Wordnet

We additionally compared the Pearson correlations between semantic distance and phonemic distance across different measures of semantic distance: (a) 1 minus the cosine

Table A3
Comparison of Pearson correlations coefficients for each word length using different semantic similarity distances (* stands for $p < .05$; ** for $p < .01$; *** for $p < .001$)

Word Length	LSA (cosine)	Wordnet (path)	Wordnet (lin)
3 letters	0.021***	0.018***	0.012***
4 letters	0.013***	0.013***	0.014***
5 letters	0.011***	0.002*	0.022***
6 letters	0.004***	0.011*	0.037*
7 letters	0.01**	0.015**	0.017*

distance between co-occurrence vectors obtained by training a LSA model on the English Wikipedia and (b) several measures relying on WordNet structure to produce a score to quantify the distance between two concepts. Table A3 shows such a comparison for the 3,702 nouns of the English phonemic lexicon using the Wordnet *path* measure (the minimum path length between two concepts in the WordNet network) and WordNet *lin* information content measure (Lin, 1998). Overall all semantic distance measures show the same qualitative pattern for every word length: There seems to be a positive correlation between semantic similarity and phonological distance in the English lexicon showing that semantically similar nouns tend also to be phonologically similar.