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

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Steven T Piantadosi

Wordform similarity increases with semantic similarity: an analysis of 101 languages

Isabelle Dautriche*[1], Kyle Mahowald[2], Edward Gibson[2], and Steven T. Piantadosi[3]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France

[2]Department of Brain and Cognitive Science, MIT

[3]Department of Brain and Cognitive Sciences, University of Rochester

May 18, 2015

Keywords: lexicon, phonetics, semantics, lexical design

Abstract

Although the mapping between form and meaning is often regarded as arbitrary, there are in fact well-known constraints on wordforms which are the result of functional pressures associated with language use and its acquisition. In particular, languages have been shown to encode some meaning distinction in their sound properties that are described to be important for language learning. Here, we investigate the relationship between semantic distance and phonological distance at the large-scale structure of the lexicon. We show evidence in 101 languages from a diverse array of language families that more semantically similar word pairs are also more phonologically

*For correspondence, e-mail isabelle.dautriche@gmail.com

similar. We argue that there is a pervasive functional advantage for lexicons to have semantically similar words be phonologically similar as well.

1 Introduction

Why do languages have the set of wordforms that they do? Although the mapping between form and meaning is often regarded as arbitrary (de Saussure, 1916; Hockett, 1960), there are in fact well established regularities in lexical systems. The simplest of these involve correlations between word length and frequency (Zipf, 1949) or informativity (Piantadosi et al., 2011). 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) (Mahowald, Dautriche, Gibson, Christophe, & Piantadosi, *submitted*), observed particularly in high-frequency words (Mahowald, Dautriche, Gibson, & Piantadosi, *submitted*). Deeply semantic regularities are are also observed: sound symbolism, in which languages encode some meaning distinction in their sound properties,[1] is one such form-meaning regularity and is present across many languages and cultures (e.g., Bremner et al., 2013; Childs, 1994; Hamano, 1998; Kim, 1977). For instance, adults intuitively pair 'bouba' with a picture of a rounded object while they pair 'kiki' with a picture of a spiky object (the "bouba-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 are present in the phonological form of words more transparently. For instance, semantic features, such as objects vs. actions, that are associated

[1] Note that this is not specific to spoken languages, sign languages do also map meanings into visual sign (see Strickland et al. *in press*)

with grammatical distinctions may be marked morphologically (Monaghan & Christiansen, 2008; Pinker, 1984).

Several studies suggest that systematic form-meaning mappings may facilitate word learning (e.g., Imai & Kita, 2014; Monaghan et al., 2011). The idea is that learning similarities among referents (and hence forming semantic categories) may be facilitated if these similarities appear also at the level of the wordform. For instance, it might be easier to learn the association of *fep* and *feb* to CAT and DOG than to CAT and UMBRELLA. This advantage in learning may be a explanation for the observation of sound-symbolism in languages and predicts that phonologically similar words would tend to be more semantically similar. In this spirit, several studies have established that it is easier to learn languages that are compressible (Kemp & Regier, 2012). For instance, in the limit, the easiest language to learn is a language that uses only one word to express all meanings. More generally, it should be easier to learn languages whose words tend to sound similar to each other, as *fep* and *feb*, because there is less phonetic material to learn, remember or produce (Gahl et al., 2012; Stemberger, 2004; Storkel et al., 2006; Storkel & Lee, 2011; Vitevitch & Sommers, 2003).

Yet there may also be a functional disadvantage for form-meaning regularities. Another 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 similar words. If someone said, "Weather forecast: ___ today and tomorrow" the missing word could plausibly have be 'sun' or 'wind', but it's unlikely to be 'boat' or 'John'. Therefore, one would also predict that semantically related words should be more distant in phonological space than semantically unrelated words, much like theories positing dispersion of phonemes in vowel space (e.g., Liljencrants & Lindblom, 1972).

In this work, we investigate the relationship between semantic distance and phonological dis-

tance. If there is a positive correlation between semantic distance and phonological distance—i.e., more similar wordforms are more semantically similar—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, there would be a pressure for dispersion for words' meanings to be more distinct relative to phonological distance, likely due to communicative pressures of confusability. Monaghan et al. (2014) previously examined the correlation between semantic distance and phonological distance in English. In this work, the authors found that phonologically similar words tend to be more semantically similar. While this result is telling, the sample of a single language does not indicate if form-meaning regularities in the lexicon are the product of functional pressures that universally apply, or historical accidents of English.

In the present work, the existence of large-scale data sets in a large number of languages makes it possible to investigate semantic and phonological relatedness across human language more generally. We use a dataset of 101 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 over each Wikipedia corpus) and orthographic similarity: Pearson correlations and a mixed model analysis to ensure that the correlation observed does not depend on a particular language family. Second, we probed the relation between semantic and phonological similarity by using a different measure looking at the interaction of semantic relatedness and the likelihood of finding a minimal pair. Finally, we also used a subset of 4 languages to assess whether the correlation between semantic and phonological similarity still hold in a set of monomorphemic words with phonemic representations. In sum, across all these languages we found that semantically similar words tend to be phonologically similar, providing large-scale, cross-linguistic evidence for phonological clustering of semantically similar words.

2 Methods

101 orthographic lexicons:

We extracted the lexicons of 101 languages from the Wikipedia database (as in Appendix A and Mahowald, Dautriche, Gibson, & Piantadosi (*submitted*)). We define as the lexicon of these language 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'), we only compared words of the same length (Section 3.3 presents a more rigorous analysis looking at semantic distance in monomorphemic words in a smaller number of languages).

3 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 et al., 1995) and Lexique for French (New et al., 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, they contained neither inflectional affixes (like plural *-s*) nor derivational affixes like *-ness*. 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 greater than 0. Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., 'bat') we included this form only once.

All three CELEX dictionaries were transformed so that diphthongs were changed into 2-character

[2]Since we calculated words' semantic distance for all pairs of words of the same length, this restriction was to limit the number of possible calculations for each language.

strings. In each lexicon, we removed a small set of words containing foreign characters. This resulted in a lexicon of 5459 words for Dutch, 6512 words for English, 4219 words for German and 6782 words for French.

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:** we used the edit distance, or Levenshtein distance between the two orthographic strings in the case of the 101 orthographic lexicons and phonemic strings in the case of the 4 phonemic lexicons. The smaller the distance, the more similar word-forms are to each other. For example, the words 'cat' and 'car' are very similar, with an edit distance of 1.

- **Semantic distance**: we used Latent Semantic Analysis (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). 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 with the same vocabulary (e.g., animal, domestic, pet, etc.). 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 on Wikipedia for each language using the `Gensim` package (Rehurek & So-

jka, 2010) in the R programming language (R Core Team, 2013). 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[3] based on (Fourtassi & Dupoux, 2013; Rehurek & Sojka, 2010). We also discarded words that appear in less 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 101 languages

3.1.1 Pearson correlations analysis

For each language, we computed Pearson correlations between the semantic distance of all pairs of words of the same length (focusing on words of length 3 to 7) and the pairs' orthographic distance.

[3]For Buginese which was the only language having less documents than 500 (the number of dimensions), we took $d = 20$ based on (Fourtassi & Dupoux, 2013).

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. 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 30 times and computed Pearson correlations for each word length. We then asked whether the correlation between wordform distance and semantic distance of the real lexicons falls outside the range of correlation values that could be expected by chance, where chance means random form-meaning assignments.

Figure 1 summarizes this hypothesis test for 4-letter words across the 101 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 but one (Buginese[4]); b) most of the correlations are significantly positive (in 74/101 languages; dark colors) meaning that the correlation between semantic distance and orthographic distance is more positive that what would be expected by chance alone.

[4]Recall that Buginese was our smaller corpora. Inspection of the words of that corpus revealed that, in addition, most of the nouns were names of places (on average 60% from a random samples of 100 words).

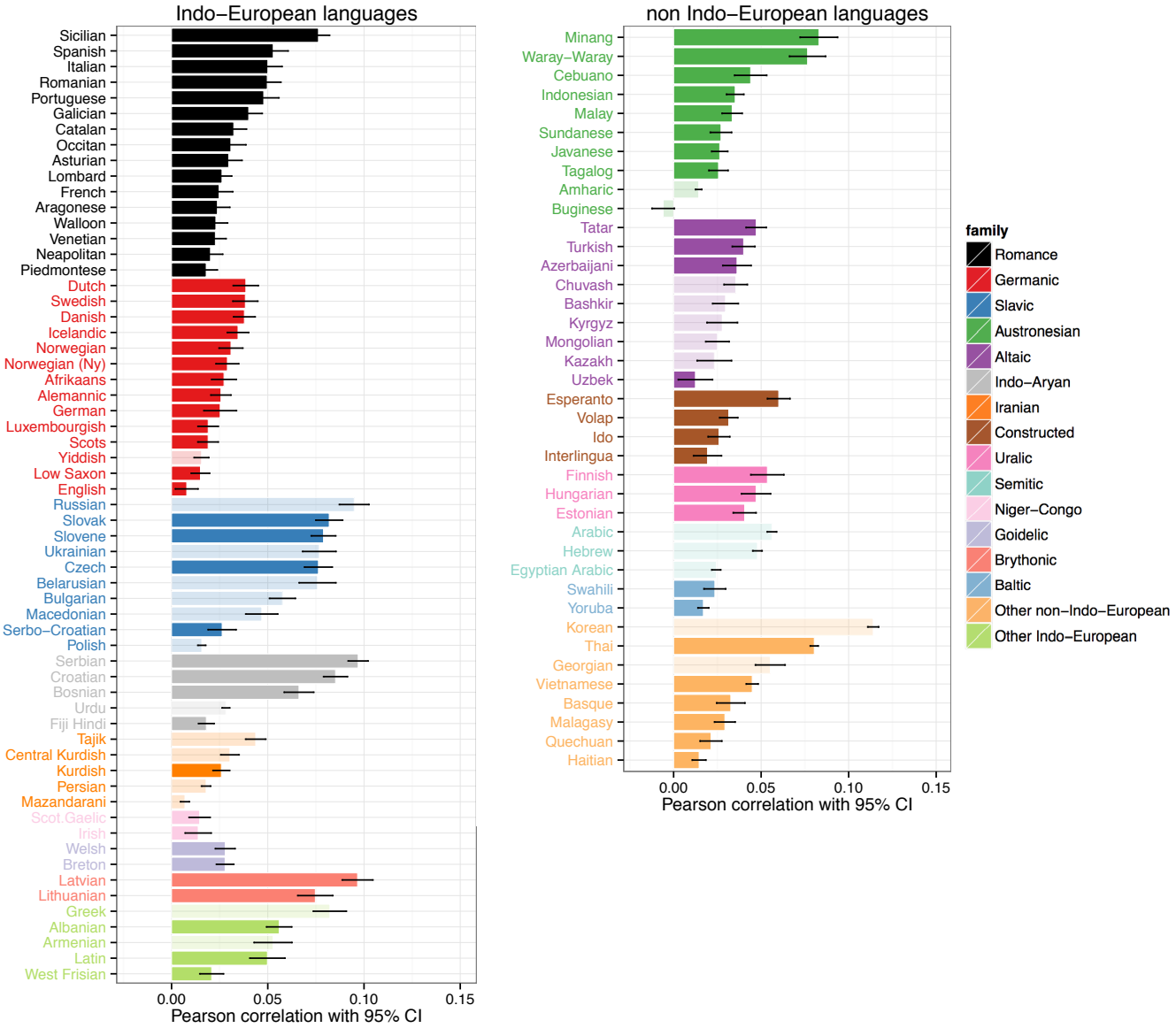

Figure 1: Pearson correlation between semantic distance (1 - cosine) and orthographic distance (Levenshtein distance) for each language for word of length 4. Languages are grouped per language family for Indo-European languages (left plot) and non Indo-European languages (right plot). Dark colors are used for significant Pearson correlations ($p < .05$) and light colors for non-significant correlations.

As in standard null hypothesis testing, we compute a z-score using the mean and standard deviation of correlations scores estimated from these 30 meanings 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. Yet, even though the correlation is highly significant, one needs to observe that this is a tiny effect explaining only a very small amount of the variance ($r < 0.05$).

word length	**mean correlation**	**proportion showing positive correlation**	**proportion showing significant correlation**
3 letters	0.049	1	0.72
4 letters	0.041	0.99	0.74
5 letters	0.040	0.99	0.73
6 letters	0.040	1	0.94
7 letters	0.047	0.91	0.71

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 $p < .05$, chance being the correlation obtained during 30 random form-meaning reassignments.

3.1.2 Mixed effect analysis

To ensure that the observed effect does not depend on a particular language family, we ran a mixed effect regression predicting scaled semantic distance for each pair of words from the Levenshtein distance between the word of the pair. We used a maximal random effect structure with random intercepts for each language, language sub-family, and language family and slopes for Levenshtein 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). 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 Levenshtein distance on semantic distance. For every word length, the coefficient for Levenshtein distance is significantly positive meaning that increased semantic distance comes with increased Levenshtein distance beyond effects of language family or sub-family.

word length	**Levenshtein distance**
3 letters	.11 ***
4 letters	.07 ***
5 letters	.06 ***
6 letters	.04 ***
7 letters	.04 ***

Table 2: Summary of the full models including random intercepts and slopes for language, sub-family, and family for Levenshtein distance for each word length. Three asterisks means that by a likelihood test, the predictor significantly improves model fit at $p < .001$.

3.2 Likelihood of finding a minimal pair in 101 languages

In addition we looked at the interaction of 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 n_{top}, and in the bottom 10% of semantically related words pairs, n_{bottom}, by looking at the ratio $\frac{n_{top}}{n_{bottom}}$. 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. Figure 2 shows the histogram of the distribution of ratio $\frac{n_{top}}{n_{bottom}}$ across all languages. As we can observe, in all 101 languages, minimal pairs are on average 3.52 (median of the distribution) more likely to appear in the top 10% semantically related words than in the least 10% related words.[5]

3.3 Generalizing form-meaning regularity to monomorphemic words

One obvious explanation for the positive correlation between semantic and orthographic distances is the presence of morphological regularity among the 101 lexicons we studied here. Even though we studied words of the same length to limit this effect, there is certainly some morphological regularity remaining (e.g., 'capitalist' / 'capitalism'). To 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 mono-morphemic codes are readily available.

For the monomorphemes of Dutch, English, French and German, we computed Pearson correlations between semantic distance and phonological distance for each word length and compared it to the correlations obtained after 30 random form-meaning reassignments. As shown in Figure 3, the

[5]Note that we obtain qualitatively the same results by looking at the 25% most related and the 25% least related words or other percentages.

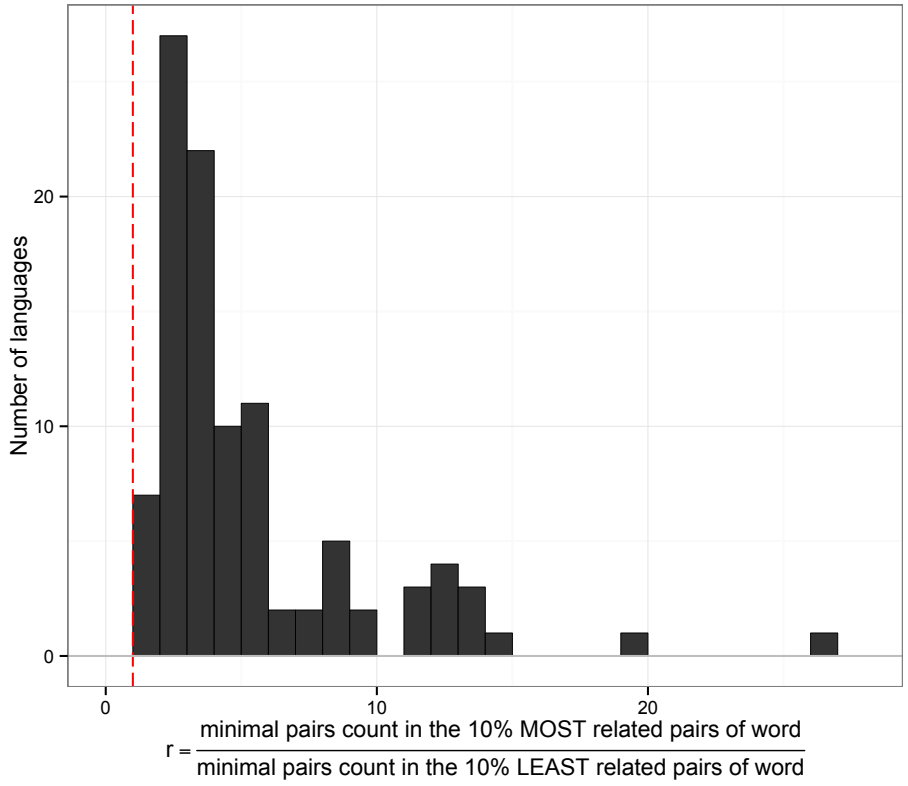

Figure 2: Distribution of the ratio of the number of minimal pairs in the 10% most related words compared to the number of minimal pairs in the 10% least related words in a given lexicon, across all the languages. A ratio below 1 means that there are more minimal pairs in semantically unrelated words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

correlations obtained in the real lexicons for each word length (the red dot) tend to be significantly more positive than the correlations obtained in 30 random configurations of form-meaning pairings (the histograms) (see also Table 3).

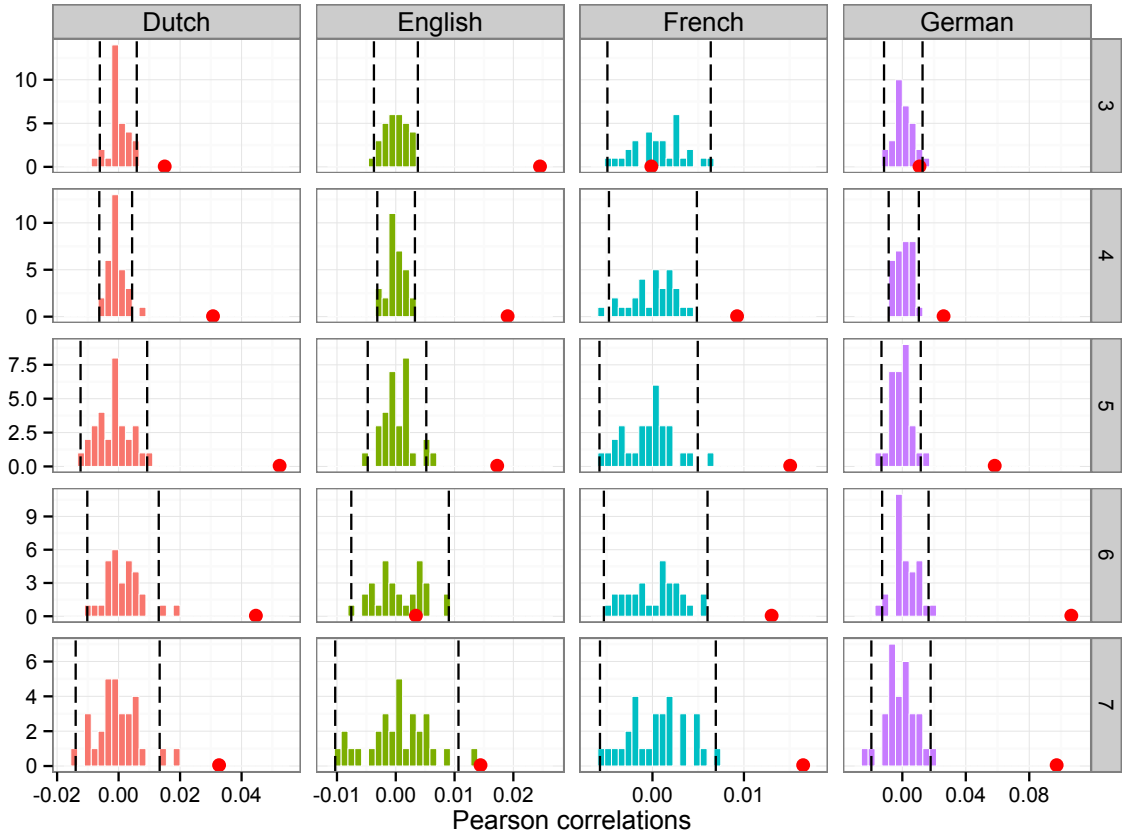

Figure 3: Pearson correlations between semantic distance and phonological distance for word of length 3 to 7 (in rows) for Dutch, English, French and German. Each histogram shows a distribution of correlations obtained after 30 random form-meaning assignments (chance level). The red dots are the correlations found in the real lexicon for that particular length. The dotted lines represent the 95% interval. The red dots tend to be to the right of the histograms.

Word length		Dutch	English	French	German
3	real	0.015	0.025	0	0.011
	r (simulated)	0	0	0.001	0.001
	σ (simulated)	0.003	0.002	0.003	0.006
	z	4.9	12.9	-0.3	1.7
	p	**<.001**	**<.001**	0.769	0.099
4	real	0.031	0.019	0.009	0.026
	r (simulated)	-0.001	0	0	0.001
	σ (simulated)	0.003	0.002	0.002	0.005
	z	11.6	11.6	3.8	5.2
	p	**<.001**	**<.001**	**<.001**	**<.001**
5	real	0.052	0.017	0.015	0.058
	r (simulated)	-0.002	0	0	-0.001
	σ (simulated)	0.006	0.003	0.003	0.006
	z	9.8	6.7	5.6	9.3
	p	**<.001**	**<.001**	**<.001**	**<.001**
6	real	0.045	0.003	0.013	0.107
	r (simulated)	0.001	0.001	0	0.002
	σ (simulated)	0.006	0.004	0.003	0.007
	z	7.3	0.6	4.4	14
	p	**<.001**	0.525	**<.001**	**<.001**
7	real	**<.05**	**<.05**	**<.05**	0.097
	r (simulated)	0	0	0.001	-0.001
	σ (simulated)	0.007	0.005	0.003	0.01
	z	4.7	2.7	4.9	10.3
	p	**<.001**	**<.01**	**<.001**	**<.001**

Table 3: z-statistics comparing the Pearson correlations (r) between semantic distance (1 - cosine) and orthographic distance (Levenshtein distance) for each word length (2 to 7 phones) and each language with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of Pearson correlations obtained in 30 random form-meaning mappings for each word length and each language (see Figure 3).

Overall semantic distance is positively correlated with phonological distance ($r = 0.04$) significantly more than what would be expected by chance ($p < .001$ across all lengths and all languages). Thus we replicate the pattern observed among the 101 lexicons: similar wordforms tend to be more

semantically similar than distinct wordforms. This is not the result of morphological similarity here since we looked only at monomorphemes in these four languages.

4 General Discussion

We have shown that across 101 languages, similar sounding words tend also to be more semantically similar above and beyond what could be expected by chance (an extension of Monaghan et al. (2014) in English). In order to remove the contribution of morphology from this correlation, we conducted the same analysis on the set of monomorphemic lemmas of a restricted number of languages and found exactly the same pattern of results. This suggests that the pattern of clumpiness in the lexicon may be in part explained by form-meaning regularities, over and beyond morphological regularity, across a large range of typologically different languages.

What could be the reasons of form-meaning regularity in the lexicon? One 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. Similarly, the presence of local sound-symbolism (e.g., the phonesthemes *gl-* in English) may drive the correlation. Yet, previous work showed that neither etymological roots nor small clusters of sound symbolic words were sufficient to account for the pattern of systematicity observed across the English lexicon (Monaghan et al., 2014). Though this needs to be confirmed for the languages under study here, this suggests a global pattern of form-meaning systematicity across the whole lexicon over and above etymological roots.

Another possibility is that form-meaning regularity is 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 Mahowald, Dautriche, Gibson, Christophe, & Piantadosi (*submitted*)), 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 (Monaghan et al., 2011) and may be one of the outcomes of language transmission and evolution (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.

Still, the prevalence of wordform similarity in the lexicon conflicts in theory with communicative efficiency. Imagine a language that displays an extreme pattern 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. Nevertheless, we observed a correlation between semantic similarity and phonological distance. Perhaps, then, semantically similar words are not as confusable as one might suspect. 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 et al. (1995); speaker Creel et al. (2008)). As a result, finer-grained contextual information may be sufficient most of the time for adults' listeners to distinguish between phonologically similar words.

To our knowledge, with 101 languages in the sample, this is the largest cross-linguistic analysis showing a correlation between semantic similarity and phonological similarity among monomorphemic words showing evidence of systematicity in form-meaning mappings beyond morphological regularity (at least for Dutch, English, French and German). Ultimately, the results here suggest a functional advantage to having lexicons in which there is a positive correlation between phonetic and

semantic similarity.

References

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

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

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

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.

Cassidy, K. W., & Kelly, M. H. (1991, June). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*(3), 348–369. Retrieved 2013-04-16, from http://linkinghub.elsevier.com/retrieve/pii/0749596X9190041H doi: 10.1016/0749-596X(91)90041-H

Childs, G. T. (1994). African ideophones. *Sound symbolism*, 178–204.

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. Retrieved 2014-08-28, from http://www.sciencedirect.com/science/article/pii/S0010027707000935

de Saussure, F. (1916). *Course in general linguistics*. Open Court Publishing Company.

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. *ACL 2013*, 165. Retrieved 2013-10-26, from http://www.aclweb.org/anthology-new/P/P13/P13-3.pdf#page=177

Gahl, S., Yao, Y., & Johnson, K. (2012, May). 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

Hamano, S. (1998). *The sound-symbolic system of japanese.* ERIC.

Harris, Z. S. (1954). Distributional structure. *Word*. Retrieved 2014-10-22, from http://psycnet.apa.org/psycinfo/1956-02807-001

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. Retrieved 2014-08-30, from http://www.pnas.org/content/105/31/10681.short

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. Retrieved 2014-10-22, from http://psycnet.apa.org/journals/rev/104/2/211/

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862. Retrieved 2015-05-07, from http://www.jstor.org/stable/411991

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML* (Vol. 98, pp. 296–304). Retrieved 2014-11-25, from http://webdocs.cs.ualberta.ca/~lindek/papers/sim.pdf

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159–165. Retrieved 2014-10-22, from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5392672

Mahowald, K., Dautriche, I., Gibson, E., Christophe, A., & Piantadosi, S. (*submitted*). Lexical clustering in efficient language design.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. (*submitted*). Cross-linguistic effects of frequency on wordform similarity.

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

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*.

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.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora.* Retrieved 2014-10-22, from http://www.muni.cz/research/publications/884893

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. (*in press*). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*.

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.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

A Appendix: Dataset of 101 lexicons from Wikipedia

We started with lexicons of 115 languages from their Wikipedia databases (`https://dumps.wikimedia.org`). We then excluded languages for which a spot-check for non-native (usually English) words in the top 100 most frequent words in the lexicon between 3 and 7 characters revealed more than 80% of words were not native. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the 3-7 letter words in Chinese Wikipedia are often English. However, we included languages like Korean in which words generally consist of several characters. After these exclusions, 101 languages remained.[6] 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 4 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 4 and 5.

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 3 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 99). The most common intrusion in these languages was English words.

[6] We excluded: Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, Kannada

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

Table 4: Table of Indo-European languages used, language families in bold.

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

Table 5: Table of non-Indo-European languages used, language families in bold.

B Appendix: 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 distance between co-occurence 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 6 shows such a comparison for the 3702 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.

word length	**LSA (cosine)**	**wordnet (path)**	**wordnet (lin)**
3 letters	.021 ***	.018 ***	.012 ***
4 letters	.013 ***	.013***	.014 ***
5 letters	.011 ***	.002 *	.022 ***
6 letters	.004 ***	.011 *	.037 *
7 letters	0.01 **	.015 **	.017 *

Table 6: Comparison of Pearson correlations coefficients for each word length using different semantic similarity distances.