On the semantics of non-words and their lexical category

Giovanni Cassani

University of Antwerp, CLiPS

Yu-Ying Chuang

Seminar für Sprachwissenschaft, Universität Tübingen

R. Harald Baayen

Seminar für Sprachwissenschaft, Universität Tübingen

Abstract

Using computational simulations, this work demonstrates that it is possible to learn a systematic relation between words' sound and their meanings. The sound-meaning relation was learned from a corpus of phonologically transcribed child-directed speech by using the Linear Discriminative Learning (LDL) framework (Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019), which implements linear mappings between words' form vectors and semantic vectors. Presented with the form vectors of 16 non-words, taken from a study on word learning (Fitneva, Christiansen, & Monaghan, 2009), the network generated the estimated semantic vectors of the non-words. As half of these non-words were created to phonologically resemble English nouns and the other half were phonologically similar to English verbs, we assessed whether the estimated semantic vectors for these non-words reflect this word category difference. In seven different simulations, Linear Discriminant Analysis (LDA) successfully discriminated between noun-like non-words and verb-like non-words, based on their semantic relation to the words in the lexicon. Furthermore, how well LDA categorized a non-word correlated well with a phonological typicality measure (i.e., the degree of its form being noun-like or verb-like) and with children's performance in an entity/action discrimination task. On the one hand, the results suggest that children can infer the implicit meaning of a word directly from its sound. On the other hand, this study shows that non-words do land in semantic space, such that children can capitalize on their semantic relations with other elements in the lexicon to decide whether a non-word is more likely to denote an entity or an action.

Keywords: Non-words; Semantics; Lexical categories; Linear discriminative learning; Phonological bootstrapping

On the semantics of non-words and their lexical category

Introduction

One of the most important tenets of modern general linguistics is that the relation between the form of a word and the meaning it refers to is arbitrary, such that it is impossible to gauge the meaning of a word solely based on its form and vice versa (de Saussure, 1916). This position is supported, among other things, by the observation that different languages use radically different forms to refer to the same concepts. For example, while in English the mammal with four paws that barks and waives its tail when happy is called *dog*, it takes different forms in other languages: *hund* (German), *cane* (Italian), *chien* (French), *perro* (Spanish), only to name a few. From the communicative and evolutionary perspectives, it is advantageous to make forms or names maximally discernible because this design helps to reduce the risk of confusability and misunderstanding (Gasser, 2004). Thus, a *wolf* would not be easily mistaken as a *dog* because of their very different names, even though their meanings, perceptual features and contexts of occurrence are similar to a substantial extent. The polarization of forms responds to the communicative pressure of differentiating two similar entities with which we however need to interact differently.

However, arbitrariness comes at the cost of learnability, since it makes the form-meaning relation unpredictable (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017; Monaghan, Christiansen, Farmer, & Fitneva, 2011). Nonetheless, it has been indicated that some parts of natural languages are not entirely arbitrary (Chater & Christiansen, 2010; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Monaghan, Christiansen, Farmer, & Fitneva, 2011; Monaghan, Mattock, & Walker, 2012; Monaghan, Shillcock, Christiansen, & Kirby, 2014). Beyond onomatopoeia, well-known examples such as sound symbolism (Hinton, Nichols, & Ohala, 1994; Imai & Kita, 2014; Maurer, Pathman, & Mondloch, 2006; Nuckolls, 1999; Sapir, 1929; Sidhu & Pexman, 2015; Westbury, Hollis, Sidhu, & Pexman, 2017) and phonaestemes (Bergen, 2004; Pastizzo & Feldman, 2009; Åsa, 1999) suggest that certain sub-lexical sound sequences are consistently mapped onto certain meanings. For

example, high vowels tend to be associated with small shapes while low vowels tend to be associated with larger shapes (size-sound symbolism, Nuckolls, 1999). Moreover, there appears to be a correlation between certain phonological traits and object shapes (D'Onofrio, 2013; Köhler, 1929; Ramachandran & Hubbard, 2001), the relation of which also affects language learning in infants (Maurer et al., 2006) (although see Styles and Gawne (2017) for a qualification of the situations in which the so-called *bouba-kiki* effect is found). Furthermore, it has recently been shown that the first phoneme of a word is a reliable predictor of the valence of the word itself, and a better predictor than subsequent phonemes (Adelman, Estes, & Cossu, 2018). This pattern was observed cross-linguistically, and reflects a pressure on communication systems to synergize crucial aspects of the world being communicated with the forms used to communicate. Thus, while arbitrariness and irregularity in languages enhance discriminability (Gasser, 2004), systematicity and regularity serve to increase predictability and learnability (Blevins, Milin, & Ramscar, 2017a; Nielsen & Rendall, 2014; Nygaard, Cook, & Namy, 2009). In this respect, the facilitatory effect of sound-symbolism for word learning has been documented by a variety of studies which targeted both children (Imai, Kita, Nagumo, & Okada, 2008; Imai et al., 2015; Kantartzis, Imai, & Kita, 2011) and adults (Lockwood, Dingemanse, & Hagoort, 2016; Monaghan et al., 2012; Nygaard et al., 2009).

Evidence that the non-arbitrary aspects of language facilitate children to learn their native languages has been reported beyond word learning. One line of research in this field considers the relation between form and grammar, suggesting that this might be rather systematic (Farmer, Christiansen, & Monaghan, 2006; Sharpe & Marantz, 2017). The main idea is that the sound of a word can provide useful and reliable cues to the lexical category of the word (Cassidy & Kelly, 1991; Durieux & Gillis, 2001; Kelly, 1988; Morgan & Demuth, 1996; Sharpe & Marantz, 2017). Children are thus hypothesized to make use of these correlations between sound and grammar to bootstrap the acquisition of lexical categories (Christophe, Guasti, Nespor, Dupoux, & Van Ooyen, 1997; Fitneva et al., 2009). This hypothesis, formulated in the context of

language acquisition, goes under the name of *phonological bootstrapping*. The central assumption is that the mapping between form and meaning is not a direct one, but an indirect one that is mediated by grammar. Therefore, when presented with a new word, real or nonce, children can use its form to gauge its grammatical category, which is in turn used to form expectations about the meaning of the new word.

The mediation of lexical categories between form and meaning complies with the general tenet of the bootstrapping theory (Gillis & Ravid, 2009), and it has been shown that the bootstrapping process can be driven by different sources of information. For the hypothesis of phonological bootstrapping as described above (Fitneva et al., 2009; Morgan & Demuth, 1996), the central assumption is that children exploit correlations between sub-segmental phonological features of words and their lexical categories to first map a word to its likely grammatical function and then use this to infer the likely meaning of the word. Syntactic bootstrapping (Gleitman, 1990) and distributional bootstrapping (Maratsos & Chalkley, 1980) hypothesize that children exploit the abstract or superficial syntactic environment of a word respectively to determine its lexical category and then use it to gauge its likely meaning. As for prosodic bootstrapping (Christophe, Millotte, Bernal, & Lidz, 2008), prosodic contours are assumed to be used to chunk the utterance and obtain a rudimentary syntactic analysis to constrain the interpretation of the likely lexical category of a word, and then use it to infer its likely meaning (de Carvalho, He, Lidz, & Christophe, 2019). A notable exception is semantic bootstrapping (Pinker, 1984), in which semantic information is assumed to directly influence the lexical category of a word, reversing the mapping. In this study, we focus on the phonological bootstrapping hypothesis, which considers the relation between word forms and their lexical categories, used as a bridge to word meaning and interpretation.

The phonological bootstrapping hypothesis has been tested empirically by Fitneva et al. (2009). Seven-year-old children were aurally presented with 16 monosyllabic non-words and asked to match those non-words to one of two pictures, one depicting an entity and the other depicting an action. The non-words were created to be

phonologically similar to English nouns or to English verbs, clustering at the two ends of the *phonological typicality* spectrum[1]. The underlying assumption of the study is that children can consistently associate noun-like non-words with pictures of entities by realizing that noun-like non-words sound more like nouns, which typically refer to entities (and the same for verb-like non-words, verbs, and actions). Therefore, according to the phonological bootstrapping hypothesis, for children to succeed in this task, they have to first map each non-word to its likely lexical category, and then decide whether to pick the action or entity picture given the inferred lexical category. Results show a positive and significant correlation between the proportion of children who chose the action referent for a word and its phonological typicality (Pearson's $r = 0.664, p < 0.01$)[2]. From this, the authors conclude that it is indeed the case that children are sensitive to the phonological form of the word, from which they infer the likely lexical category to gauge the word's meaning.

In the present study, we test the hypothesis that the relation between form and meaning is not necessarily mediated by grammar, in the sense that abstract lexical categories are crucially involved, and that instead there is a non-arbitrary mapping directly linking form and meaning (Dingemanse et al., 2015; Nygaard et al., 2009). This mapping is hypothesized to generalize to completely novel forms, including non-words. Using the stimuli from Fitneva et al. (2009), we explore whether the same distinction between noun-like and verb-like non-words which characterizes their phonology also

---

[1] For a detailed description of how phonological typicality is defined and computed see (Farmer et al., 2006; Monaghan, Christiansen, & Fitneva, 2011).

[2] The correlation was computed from the original data set by only considering trials in which children could exclusively rely on the sound of a word to decide between the entity or action picture. The proportion of children who picked the action referent was chosen as the dependent variable since phonological typicality is computed in such a way that noun-like non-words have negative scores and verb-like non-words receive positive scores: using the proportion of children who picked the entity referent would result in a correlation with the same magnitude but an opposite sign. The interpretation, however, would not change: the more a non-word sounds like a verb (i.e. the higher its phonological typicality score), the more the children who picked the action referent for it, and vice versa.

reflects onto their semantics. This would entail that children may have inferred the likely referent of each non-word straight from their sounds, bypassing the intermediate step of lexical category identification. It is important to note that the target non-words were not created to incorporate phonaestames, sound-symbolic features, or morpho-phonology: therefore, if our hypothesis is correct, our study would demonstrate a stronger systematicity in the relation between form and meaning than previously assumed, to the point that whatever form could in principle generate a semantic impression, the reliability of which depends on the form-to-meaning mappings already in the lexicon. Unlike the phonological bootstrapping hypothesis, in which children are first assumed to map the phonological form to the likely lexical category, and then use the lexical category to constrain meaning interpretation, we hypothesize that children can immediately exploit semantic relations evoked from phonology, and that even isolated non-words actually also elicit an informative semantic impression.

The direct mapping from form onto meaning examined in this study is obtained by implementing Linear Discriminative Learning (LDL), a computational model of the mental lexicon put forward by Baayen et al. (2019). Mappings between form and meaning are estimated using standard linear transformations from linear algebra, which are equivalent to two-layer networks without any hidden layers. Given the form-to-meaning mapping, inputting the form vector of a word will return the semantic vector of the word. It has been shown that with LDL, high accuracies can be achieved for both visual and auditory single word recognition (Baayen et al., 2019). The same mapping can be used to generate the semantic vectors of non-words as well. Given the form vector of a non-word, the network can estimate its meaning and return the predicted semantic vector of the non-word. The non-word, therefore, is assumed to exist in the same semantic space as other lexical or sub-lexical (e.g., inflectional functions) elements in the lexicon, and how this non-word is semantically related to these elements can then be measured.

In total seven simulations were conducted in this study to assess whether a systematic form-to-meaning mapping exists, so that the semantic clustering of

non-words can directly reflect the phonological clustering, rendering superfluous mediation by lexical categories. The first three simulations make use of three different sources of semantic information, in which the degree of the involvement of lexical categories gradually decreases. For the first simulation, the lexical categories of each non-word's semantic neighbors (semantically similar words) are considered. Here we test the hypothesis that it is indeed possible to derive a mapping from form to lexical categories, which can be used by children to infer the meaning of non-words. Importantly, the mapping between form and lexical categories relies on semantic information (in line with the semantic bootstrapping hypothesis), since the neighbors whose lexical category is considered are defined on semantic grounds rather than on purely formal grounds. This approach is different from the phonological bootstrapping hypothesis (Christophe et al., 1997; Fitneva et al., 2009), where the link between form and meaning is indirect while the sound maps directly onto lexical category information. However, it shares with the phonological bootstrapping hypothesis the intuition that word category information is driving children's semantic responses.

The second simulation goes one step further by examining the relation between non-word meanings and the meanings of highly grammaticalized elements such as morpho-syntactic functions (e.g., PLURAL or PAST). Information about lexical categories can be implicitly encoded in morpho-syntactic functions[3]. It should be emphasized that no lexical categories are explicitly built into the model. Instead, following Westbury and Hollis (2018), we assume that lexical categories can emerge from the distribution of lexical items, and need not be explicitly modeled. The second simulation, therefore, tests whether children can do without the mediation of lexical categories and still infer the intended clustering by relying on the semantic content of morpho-syntactic functions. In the third simulation, we take a radical approach by comparing the meaning of non-words with the meaning of prototypical and developmentally salient words for things (e.g., *ball*) and actions (e.g., *cry*). By doing so, we investigate whether

[3] For example, PAST characterizes verbs and SUPERLATIVE is a property of adjectives, although number, by contrast, can be expressed on both nouns and verbs

the children in the experiment by Fitneva et al. (2009) could have inferred the *entity-ness/action-ness* of a non-word simply from its semantic relation to a few developmentally salient words, without the need to resort to lexical categories.

As the information contributed by the three semantic relations described above is not mutually exclusive and children might make use of all of them, possibly to different extents, to make an *entity/action* judgment, for the rest of the simulations, different combinations of the three sources of semantic information are considered, as an attempt to estimate the importance of each source. Thus, the fourth simulation combines neighbors' lexical categories and morpho-syntactic functions, the fifth simulation combines neighbors' lexical categories with anchor words, and the sixth simulation combines morpho-syntactic functions with anchor words. For the last simulation, all three sources of semantic information are considered together.

To sum up, this study examines whether the inference of lexical category from the sound pattern of a newly encountered word (hence a non-word) is an indispensable intermediate step that bootstraps word learning, the assumption underlying the phonological bootstrapping hypothesis which was empirically tested by Fitneva et al. (2009). Taking a different approach, we explore the possibility that the semantic content of the non-word and its semantic relations with other elements in the lexicon are what drive the behavioral patterns observed in Fitneva et al. (2009). It is worth mentioning that within our approach, lexical categories are conceptualized as graded constructs. It is not assumed that category membership is binary, with words that can either be members of a lexical category or not. On the contrary, words can reflect a broader category to different degrees (Sharpe, Reddigari, Pylkkänen, & Marantz, 2018; Westbury & Hollis, 2018). More radically, categories do not need to exist as independent constructs for this account to work (Ambridge, 2017; Ramscar & Port, 2015): the semantic vector generated for each non-word only needs to enter an informative relation with the words in the lexicon, which can be operationalized using the concepts of distance and similarity (Goldstone, 1994; Sloutsky, 2003). Observed categorical behavior is considered to be an emergent property of the system, which

manifests itself in the behavioral response required by the experimental paradigm and is made possible by the information encoded in the relation between the semantic impression evoked by non-words and the semantic knowledge in the mental lexicon. The focus of our study is the exploration of this type of information and its reliability in capturing relations which could give rise to categorically constrained behavior even in the absence of abstract categories.

Methods

Stimuli

16 non-words from the study by Fitneva et al. (2009) were used as stimuli. Eight of them were created to phonologically resemble English nouns, while the other eight were created to sound like English verbs[4]. Phonological typicality was computed from the phonological transcriptions retrieved from CELEX, following the method described in Farmer et al. (2006) and Monaghan, Christiansen, and Fitneva (2011), which relies on average Euclidean distance of phonological feature matrices to estimate the degree of phonological similarity between a non-word and all verbs and nouns separately. In further detail, words and non-words are represented using phonological features. Then, for each target non-word a noun typicality and a verb typicality score are calculated by computing the average Euclidean distance between the phonology of the target non-word and that of all monosyllabic, mono-morphemic nouns and verbs in CELEX. The noun typicality is then subtracted from the verb typicality, yielding a summary score which indicates whether a target non-word is phonologically closer to the average noun or verb (see the original study by Fitneva et al. (2009) for the details of the procedure used to create the target non-words). We phonologically encoded the non-words according to the phonological transcription provided in CELEX and used the transcribed forms as input to our computational model.

---

[4] The 16 non-words were the following: hæps, gælv, mɛfs, pɹælt, pɔsp, lɔfs, ɹæf, ɹɪsp, fɛlg, dwɪg, skɪk, stɔŋk, pɹʌŋ, zɪm, sɪg, smɪŋ. The first eight non-words are noun-like, the last eight words are verb-like in terms of phonological typicality

Corpus

Corpora of child-directed speech were downloaded from the British and American sections of the CHILDES database (MacWhinney, 2000) and were concatenated, preserving the chronological order of each transcript based on the age of the target child. First, the phonological form of each word in the corpus was retrieved from the CELEX database. Whenever an utterance contained a word which was not found in CELEX, the whole utterance was discarded. However, the 25 most frequent words found in the corpora but not in CELEX were hard-coded to improve coverage[5]. Some of these words are not available in CELEX with the spelling found in some of the CHILDES transcripts because of American/British English variants, e.g. *color* or *favorite* or because of non standard spellings, such as *doggie* and *horsie*. Moreover, some very frequent compounds were transcribed as one word in CHILDES whereas CELEX lists them with a space in between, e.g. *byebye* and *hotdog*. Other forms such as *will'nt* or *ssh* are specific transcription choices to render contractions and interjections found in spoken language. Finally, the word *lego* is a good example of a domain specific word which is not found in CELEX. About 87% of the utterances in the corpus could be entirely recoded phonologically, resulting in a corpus of more than 1.5 M utterances, and more than 6.2 M tokens.

Then, the corpus was processed to look for possible compounds: whenever two adjacent words in the corpus were found as a possible compound in CELEX, they were joined and treated as a single word-form. Finally, the corpus was processed using the Tree Tagger (Schmid, 1994) to have a more fine-grained set of Part-of-Speech tags that allowed us to extract morpho-syntactic functions from the words in the corpus. For example, the token *wolves* was recoded as [*wolf*, PLURAL], the token *spoke* as [*speak*, PAST], and the token *wonderful* as [*wonderful*, -FUL] to indicate its derivational affix. Inflected words were represented using the base form and the affix, to highlight that the

[5] This threshold allowed to improve coverage while keeping manual work to the minimum. In order to get further sizeable increases in coverage many more words would need to be hard-coded, since their frequency in the corpus decreases quickly.

affix does not change the semantics of the word. On the contrary, derived words were represented as the whole derived form and the affix, to stress that derived words have their own semantics, which differs from that of the original form. For detailed tagging procedures, please refer to Baayen et al. (2019) and Baayen, Chuang, and Blevins (2018). This encoding is motivated by the broader theoretical framework in which this study is situated, i.e. the LDL framework. Its goal is not limited to addressing comprehension, but extends to production as well. To this end, semantic vectors for inflectional functions are essential.

The semantic vectors of non-words

We implemented the model of LDL (Baayen et al., 2019) to generate the semantic vectors of the non-words. To achieve this, a mapping that encodes the form-to-meaning relation needs to be learned from real words first. This was done by making use of the CHILDES corpus. To learn the mapping, we created two matrices: a form matrix $\boldsymbol{C_w}$ and a semantic matrix $\boldsymbol{S_w}$. $\boldsymbol{C_w}$ is a 17826 × 11722 matrix, with rows listing all the word types in CHILDES and columns indicating the tri-phones (i.e., sequences of three phones) found in all the words. In $\boldsymbol{C_w}$, the form vector of each word ($\boldsymbol{c_w}$) specifies which tri-phones are present in the word, using binary coding with 1 for presence and 0 for absence. For example, the word *sweet*, phonologically transcribed as /swit/, has four tri-phones: /#sw/, /swi/, /wit/, and /it#/, where # denotes a word boundary. Thus, the $\boldsymbol{c_w}$ of *sweet* has the value 1 for these four tri-phones and 0 for the rest of the tri-phones. The meaning matrix $\boldsymbol{S_w}$ has the same number of rows as $\boldsymbol{C_w}$, and each row represents the semantic vector ($\boldsymbol{s_w}$) of the word, which was obtained by training a lexome-to-lexome Naïve Discriminative Learning network on the CHILDES corpus. (For more details of this learning network, see Baayen, Milin, and Ramscar (2016) and Milin, Feldman, Ramscar, Hendrix, and Baayen (2017)). The original length of the semantic vector $\boldsymbol{s_w}$ (and hence the column number of $\boldsymbol{S_w}$) is 12537. However, given that a large number of column units are contributing little information to discriminating meanings due to their very low variances, we removed those semantic elements (column vectors)

with low variance[6], preserving 2168 columns in $\boldsymbol{S}$. All dimensions whose variance was lower than 1×10^{-8} were removed.

The mapping $\boldsymbol{F}$ was then obtained by multiplying the Moore-Penrose generalized inverse of $\boldsymbol{C_w}$ with $\boldsymbol{S_w}$. Since multiplying $\boldsymbol{C_w}$ with $\boldsymbol{F}$ gives us $\boldsymbol{S_w}$, as in:

$$\boldsymbol{C_w F} = \boldsymbol{S_w}, \tag{1}$$

the estimated semantic matrix of non-words ($\hat{\boldsymbol{S}}_{nw}$) can be generated by multiplying the form matrix of non-words ($\boldsymbol{C_{nw}}$) with $\boldsymbol{F}$:

$$\boldsymbol{C_{nw} F} = \hat{\boldsymbol{S}}_{nw}, \tag{2}$$

see Baayen et al. (2019) for further mathematical details.

$\boldsymbol{C_{nw}}$ is a 16×11722 matrix: each of its rows is the binary-coded form vector for one of the 16 non-words in Fitneva et al. (2009). However, given that the tri-phone list of the original data set did not contain all the tri-phones for these non-words, newly-occurred tri-phones had zero weights in the mapping $\boldsymbol{F}$, and as a consequence did not effectively contribute to the semantic vectors estimated for non-words. Nine out of the 16 non-words had all their tri-phones covered in $\boldsymbol{C_w}$. Six of them missed one

[6] Dimensionality reduction is necessary in order to eliminate columns which mostly consist of 0s and thus do not contribute (almost) any information but increase the computational burden. The decision of simply trimming low-variance dimensions follows from the underlying principles of the LDL framework, which aims to keep its constructs as linguistically transparent as possible. Therefore, trimming was preferred to other methods such as Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF) and others, which project the original multi-dimensional space on a lower-dimensional space which preserves the largest amounts of variance, and have been widely used in semantic models (Landauer & Dumais, 1997; Lund & Burgess, 1996). In SVD or NMF the dimensions of the smaller space are opaque, since they typically conflate several original dimensions, creating latent dimensions whose interpretation is difficult and unclear. By simply trimming low-variance dimensions, on the contrary, it is possible to know exactly which words contribute to the definition of the semantic space. The residual collinearity across dimensions, which would be minimized by using SVD, NMF or similar approaches, does not represent a problem for the matrix inversion operation which is critical to map form onto meaning.

tri-phone, and only one non-word missed two. Using equation (2) we obtained $\hat{\boldsymbol{S}}_{nw}$; each row in $\hat{\boldsymbol{S}}_{nw}$ is the estimated semantic vector $\hat{\boldsymbol{s}}_{nw}$ for the corresponding non-word.

Sources of information

As previously mentioned in the introduction, we evaluate three different sources of semantic information which could play a role in explaining the behavior observed by Fitneva et al. (2009), where children's referential choices were influenced by the phonological typicality of isolated non-words. The first of the three sources of information consists of the distribution of lexical categories across the semantic neighbors of a non-word. The second consists of the relation between the estimated semantic vector of a non-word and the semantics of morpho-syntactic functions. The third consists of the relation between the estimated semantic vector of a non-word and the semantics of a handful of early acquired, prototypical nouns and verbs. These three sources of information and their operationalization are discussed in the following paragraphs.

The lexical categories of semantic neighbors. To account for the data in Fitneva et al. (2009), phonological bootstrapping (Christophe et al., 1997) proposes that children learn a form-to-grammar mapping that allows them to infer the likely lexical category of a word from its form, and then decide whether the word more likely denotes an action or an entity based on the correlations between lexical categories and semantics. Instead of assuming that children map the phonology of a newly encountered word to its likely lexical category, we hypothesize that children can form a semantic impression from the sound of a non-word, captured by its estimated semantic vector, and that they consider its semantic neighborhood to decide whether the non-word is more likely to refer to an entity or to an action by considering the lexical category of the neighbors. This approach is reminiscent of the semantic bootstrapping hypothesis (Pinker, 1984), as it assumes that semantic information drives lexical categorization. However, it borrows from phonological bootstrapping (Christophe et al., 1997) the idea that phonological information also influences lexical categorization, although indirectly,

by first evoking a certain semantic content.

Specifically, the estimated semantic vector of each non-word ($\hat{\boldsymbol{s}}_{nw}$) was correlated with each word's semantic vector ($\boldsymbol{s_w}$) in the trained lexicon ($\boldsymbol{S_w}$), excluding morpho-syntactic functions. The 50 words with the strongest positive correlation with each estimated semantic vector were regarded as semantic neighbors. The lexical categories with which the 50 words are tagged in CELEX were retrieved. Whenever a semantic neighbor is found with more than one lexical category, all categories are considered. We considered nine lexical categories, including nouns, verbs, adjectives, adverbs, conjunctions, determiners, pronouns, prepositions, and quantifiers as found in CELEX.

A number of implementation choices here need to be further discussed. First, correlations were used to compute the similarity across semantic vectors, since from a mathematical perspective it is analogous to computing cosine similarity on centered vectors, with cosine being a standard approach to similarity in distributional semantics. Moreover, across a variety of simulations we get very similar results (typically, cosine and correlation similarities are strongly correlated, $r = 0.99$). Moreover, the choice of how many neighbors to consider was not subject to any grid-search or optimization[7]. Finally, the threshold on which neighbors are considered is both similarity- and rank-based, as nearest neighbors are retained only if their correlation is positive (hence, higher than 0) and falls among the 50 strongest correlations. We set up the selection procedure in such a way that if fewer than 50 neighbors had a positive correlation with a target semantic vector, then only those neighbors with a positive correlation would be retained, regardless of their number. However, all target non-words had at least 100 nearest neighbors with a positive correlation, and 100 is the highest number of neighbors we consider. Finally, neighbors were not weighted by their correlation

[7] We ran two further simulations considering 25 and 100 neighbors. Patterns were remarkably similar, although considering more neighbors makes the signal stronger and easier for the lda to pick up. Considering 50 nearest neighbors strikes a balance between having enough neighbors and not having too many. In principle, any neighborhood size could be picked, but we observe that the added value of more neighbors, decreases when the neighborhood size increases.

similarity, as weighting would introduce a further degree of freedom and increase the complexity of the approach, so we preferred to avoid it.

We assume that the estimated semantic vectors for noun-like non-words land in a place of the semantic space which is populated by words with a different lexical category from that of the words which populate the neighborhoods of verb-like non-words. Meaning is not directly at stake, as each neighbor is reduced to its lexical category. Each estimated semantic vector for noun-like non-words could relate differently to the words in the lexicon and have a different semantic neighborhood. Therefore, when children have to decide between the entity picture or the action picture, they may be hypothesized to sample the closest neighbors to the estimated semantic vector given a non-word and consider their lexical categories. An alternative choice could be that of creating prototype vectors for nouns and verbs, by averaging the semantic vectors of several words from a same lexical category, following Westbury and Hollis (2018). However, in line with traditional kNN approach used in cognitive science (Nosofsky, 1990) and computational linguistics (Daelemans & van den Bosch, 2005) we prefer to base our approach on similarity to exemplars rather than to prototypes. In other words, this approach does not rely on the semantic vectors of the neighbors directly to estimate categories, as was done by Westbury and Hollis (2018), but relies on an intermediate abstraction, which extracts lexical categories from the neighborhood and considers their frequency of occurrence to find the best discrimination function separating semantic vectors estimated from noun-like non-words from semantic vectors estimated from verb-like non-words.

Morpho-syntactic functions. The second source of semantic information we consider is the relation, quantified by using correlation similarity, between each estimated semantic vector ($\hat{\boldsymbol{s}}_{nw}$) and the semantic vectors of a number of morpho-syntactic functions ($\boldsymbol{s}_{aff}$). As previously mentioned, semantic vectors were also learned for morpho-syntactic functions such as -MENT, -OUS, 3RD-PERSON, PAST, and so on: there were in total 43 different functions identified in the corpus. Our hypothesis is that the correlation between the estimated semantic vectors of non-words and the

semantic vectors of the morpho-syntactic functions preserves the phonological distinction between noun-like and verb-like non-words, which can explain the choices children made in selecting the entity or action picture when presented with a non-word. Although traces of lexical category information still remain in morpho-syntactic functions (given that some functions are solely or predominantly used for one particular lexical category), the information about lexical category conveyed by morpho-syntactic functions is reduced and more implicit than that conveyed by the lexical categories of the semantic neighbors.

Anchor words. Finally, the third source of information dispenses with grammatical information altogether by directly considering the relation, again quantified by using correlation similarity, between the estimated semantic vectors of the non-words with the semantic vectors of a few prototypical and developmentally salient nouns and verbs (*anchor words*), following the approach proposed by Westbury (2014). The *anchor words* were selected from the Age-of-Acquisition (AoA) norms collected for 30,000 English words by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). The list of *anchor words* consists of ten nouns and ten verbs, which cover different semantic domains[8]. We selected the earliest acquired 150 words and checked the lexical category with which each word was tagged in the corpus. If a word was overwhelmingly tagged as the intended category, we kept it (all words are found occurring tagged with the intended category more than 90% of the times they occur). Therefore, a word such as *smile* was discarded since it can equally be a noun or a verb in the corpus we used (44% of its occurrences were tagged as noun, 56% as verb); on the contrary, a word such as *mom* was retained because it can only be a noun (100% of its occurrences in the corpus are tagged as noun). The criteria for inclusion were the following: words in the AoA norms were considered from the earliest to the latest acquired. Words were checked for their lexical category, only considering those which could be tagged as nouns or verbs.

---

[8] Nouns include *mom, spoon, daddy, carrot, dog, bed, ball, nose, flower, house*, and verbs include *cry, eat, ask, see, go, say, try, get, sing, sit*. Words such as *cry* can be both verbs and nouns: however, in the corpus we used, they appeared more often as verbs, hence we kept them.

Words with unambiguous lexical category or with an uneven distribution of part-of-speech tags in the corpus were retained. The log-frequency of the anchor words in the corpus ranges between 2.95 and 5.01, with nouns generally having a lower frequency than verbs. Finally, all words were produced by more than 70% of the children at 30 months of age in the MacArthur-Bates Communicative Development Inventory for English (Fenson et al., 2007) (range between 73% for *say* and 99% for *ball, daddy, mom*), confirming that they are early acquired and salient to children.

The underlying assumption of this approach is that words occurring together or in similar contexts tend to be semantically similar (Firth, 1957). Thus, a non-word could be judged to be more noun-like because it sounds more like a noun in the sense that its estimated semantic vector is closer to well-learned and well-established semantic vectors for nouns than to those for verbs, rather than because of correlations between sound and lexical category. Crucially, the information pertaining to lexical categories is never explicitly considered by the LDL. Estimated semantic vectors are directly compared to the semantic vectors of the *anchor words*, with all their meaning differences. Moreover, differently from the approach taken when considering neighbors' lexical categories, target non-words are compared to the same set of *anchor words* rather than to their immediate semantic neighborhood. All information pertaining to lexical categories is implicit in the semantic vectors of the *anchor words* — which was estimated without considering lexical category information at any level — and in their relation with the estimated semantic vectors. Each *anchor noun* is different, there is no abstract label signaling that they are similar and different from *anchor verbs*: categorization can only leverage the pertinent semantic relation.

The experimental design is summarized in Figure 1. From the CHILDES corpus we derived the matrix of triphones for each word and the semantic vectors, by means of the Naïve Discriminative Learning model. Then, using the LDL framework, we got the transformation matrix from triphones to semantic vectors, which was then used to generate semantic vectors for the target non-words. The generated semantic vectors were then correlated with the observed semantic vectors, computed from the CHILDES

corpus. At this stage, three different matrices were created: the first (T1a, in the figure) provided the nearest semantic vectors for each generated semantic vector. This matrix was further trasformed by counting how many neighbors had a certain PoS tags in CELEX. The resulting matrix (T1b) was then used as input for the LDA. The second matrix (T2) provides the correlation between generated semantic vectors and the semantic vectors of morpho-syntactic functions. The third matrix (T3) provides the correlation between generated semantic vectors and the semantic vectors of 20 early-acquired, prototypical words, ten nouns and ten verbs. Both T2 and T3 were fed as input to the LDA. We then performed seven simulations: the first only considered T1b, the second only considered T2, the third considered T3, the fourth combined T1b and T2 (after having z-standardized all values in each matrix separately so to have all values on the same scale), the fifth combined T1b and T3 (again performing z-transformation), the sixth combined T2 and T3, the seventh combined T1b, T2, and T3 (again, applying a z-transformation to all values in each matrix separately.

Simulation results

The analysis is divided into two main parts. We first investigated to what extent non-words could be clustered into noun- and verb-like groups using the semantic relations outlined previously, considering the clustering accuracy and the discriminability introduced by the input dimensions. Linear Discriminant Analyses (LDA) were performed on the correlation values for lexical category classification[9], using the *lda* function from the R package *MASS* (Venables & Ripley, 2002). For the second part of the analyses, we then evaluate how much the categorization results of LDA reflect phonological typicality and behavioral patterns. The LDA log-odds values, which express the confidence of the LDA in clustering each non-word as either noun- or verb-like, were correlated with phonological typicality to verify to what extent the

[9] To ensure that any observed result was not due to the chosen classification algorithm, we replicated the analysis using Support Vector Machines (SVMs), using the *svm* function from the *e1071* package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017), since SVMs implement a very different approach to classification. Results were consistent using SVMs and LDA.

semantic information preserves the phonological distinction. The LDA log-odds were also correlated with children's behavioral data taken from Fitneva et al. (2009) to assess whether semantic relations can explain behavior in this task.

Clustering and Discriminability

As previously described, seven different simulations were carried out. First, we considered each source of semantic information individually (`neighbors`, `affixes`, `anchors`). Next, we evaluated pairs of information in combination (`neighbors + affixes`, `neighbors + anchors`, `anchors + affixes`). Finally, we put together all three sources of information (`all`). For each simulation, the following statistics are provided: first, the clustering accuracy by LDA, indicating the degree to which noun-like and verb-like non-words were clustered together with other noun-like and verb-like non-words, respectively. The second statistic of interest is Wilks' λ, a measure of the degree to which the input dimension fed to LDA is informative with respect to the target classification. Wilks' λ ranges between 0 and 1, with the lower bound indicating perfectly reliable separation and the upper bound indicating no separation at all given the chosen input dimensions. Wilks' λ was calculated using the *Wilks.lambda* function from the *rrcov* package (Todorov & Filzmoser, 2009).

Before describing the results, two methodological issues have to be acknowledged and described. First, the input to LDA differs between the simulation of `neighbor` and those of `affixes` and `anchors`. For the former, LDA receives frequency counts over lexical categories while for the latter, LDA receives correlation values. This is a problem when combining different sources of information because frequency counts and correlation values exist on different numerical scales. Therefore, in order to make the three sources of information directly comparable, we z-standardized the scores independently for each source of information.

A further problem concerns the collinearity across dimensions fed to LDA. For most of our simulations, the dimensions fed as input to LDA were collinear, suggesting that the generated semantic vectors relate similarly to two or more dimensions, be these

anchor words or morpho-syntactic functions. The *lda* function generates warning whenever column vectors in the input matrix are highly collinear. We therefore reduced the number of dimensions by successively removing columns with close to zero variance until no warnings were generated. The threshold was set so to get rid of the lowest number of dimensions while avoiding risks of adverse effects of collinearity. The information about the dimensionality of the input space for each simulation and the threshold value chosen to avoid collinearity issues is provided in the Appendix.

Table 1 provides several statistics which characterize the degree to which the LDA could use the information provided in each simulation to correctly cluster noun- and verb-like non-words in the appropriate way. These statistics include the classification accuracy, complete with an indication of which non-words were incorrectly clustered with phonologically dissimilar words, and Wilks' λ, together with information about the degrees of freedom, χ^2 statistic and corresponding p-value. First, for the three individual simulations (top panel), we can observe that clustering is effective in all cases. The accuracy of `affixes` is the lowest, but it is still significantly higher than chance. No two simulations incorrectly cluster the same target non-word, suggesting that there is no non-word which is inherently difficult, but that each source of information captures different relations between the non-words and the lexicon. As far as the Wilks' λ is concerned, all values indicate a good discrimination, ranging between 0.409 (`neighbor`) to 0.281 (`affixes`), although in all cases it is impossible to rule out that the discriminability brought by the input dimensions does not result from chance (all p-values are higher than 0.05). The reason is twofold: on the one hand, the small sample size makes it harder to rule out that observed patterns, albeit strong, do not result from chance; on the other hand, the high dimensionality of the input space with respect to the sample size enforces caution in concluding too much from the available evidence.

When two or three sources of information are combined, accuracies are always perfect, suggesting that the three sources of semantic relations might have encoded complementary information. The Wilks' λ allows us to identify subtler differences. One thing to note is that all simulations have the same number of degrees of freedom, 14,

which makes them easily comparable. Discriminability is again strong, particularly when `affixes` is part of the considered semantic information. However, discriminability is still not reliably different from chance when the lexical categories of the semantic neighbors are combined with either morpho-syntactic functions or anchor words (`neighbors + affixes`). The only two cases in which Wilks' λ reliably indicates that the input dimensions can be used to partition the target items according to the desired clustering include the combination of morpho-syntactic functions and anchor words (`anchors + affixes` and `all`). Taken together, it seems to be the case that the correlation similarity between estimated semantic vectors on the one hand and the semantic information captured by anchor words and morpho-syntactic functions captures reliable information about the implicit lexical category of the target non-words. Moreover, their informativity surpasses that of the frequency distribution of lexical categories in the semantic neighborhood of an estimated semantic vector. While this does not appear when considering the clustering accuracy, it becomes apparent when considering the robustness and reliability of the discriminability brought by the different input dimensions and combinations thereof. Once the LDA receives information about how estimated semantic vectors correlate with anchor words and morpho-syntactic functions, having information about the lexical categories of semantic neighbors' does not add further information.

Correlations

The correlation analyses focus on the degree of which the LDA log-odds are correlated with phonological typicality on the one hand and children's behavioral patterns on the other. LDA log-odds capture the certainty of the LDA in clustering target non-words as either noun-like or verb-like based on their semantic information inferred from their phonological form: more positive values indicate stronger certainty that a certain non-word is verb-like and vice versa. In an analogous way, higher values on the typicality spectrum indicate that a word is more verb-like. Behavioral patterns are quantified using the proportion of children who chose the action referent for a given

non-word, to preserve the directionality of both LDA log-odds and phonological typicality.

We checked whether correlations were significant using the *rcorr* function from the *Hmisc* package in R (Harrell Jr, with contributions from Charles Dupont, & many others., 2017). Correlations with phonological typicality and behavioral patterns obtained in different simulations were further compared across different simulations using the *cocor* function from the homonymous package (Diedenhofen & Musch, 2015), in order to verify whether the use of a certain source of information proves to be more useful than others. Finally, analyses using random forests were performed to verify which individual dimensions provided the most information to the clustering task. To do this, the *randomForest* function from the homonymous R package (Liaw & Wiener, 2002) was used, and the mean decrease in the Gini purity brought by each dimension to identify the most useful ones was considered.

The correlation data for each simulation is summarized in Table 2. In general, we observe that correlations between LDA log-odds and phonological typicality are very strong, ranging between 0.734 (`neighbors`) and 0.984 (`all`). Similar to the measure of discriminability (Table 1), this correlation is weakest when the lexical categories of the semantic neighbors are considered, and strongest when all three sources of information are combined. Nonetheless, even the weakest correlation is strong and reliable, suggesting that phonological typicality is preserved in the categorical organization of the semantic space into which non-words are projected.

The comparison of different sources of information enables us to assess which source contributes the most information. To this end, the correlations between LDA log-odds and phonological typicality obtained for the three sources of information were compared. However, no significant differences were observed.

We repeated this procedure, comparing each simulation combining two sources of information to the two simulations considering each of the two individually. Results of these comparisons are summarized in Table 3. The first panel, for example, shows the comparison between the simulation of `neighbors + affixes` with that of `neighbors`

and with that of `affixes`. Significant differences were observed for the following comparisons, using a one-tailed test under the hypothesis that the more complex simulation shows a higher correlation between LDA log-odds and typicality than the simpler simulation. Bonferroni correction was applied to take multiple comparisons into account. The `neighbors + affixes` simulation has a significantly stronger correlation with phonological typicality than the `neighbors` simulation (Hotelling's t test: $t = 4.248, df = 13, p < 0.001$) and the `affixes` simulation (Hotelling's t test: $t = 2.31, df = 13, p < 0.05$). On the contrary, the `neighbors + anchors` simulation correlates significantly more strongly with phonological typicality than the `anchors` simulation (Hotelling's t test: $t = 2.603, df = 13, p < 0.05$), but is indistinguishable from the `neighbors` simulation (Hotelling's t test: $t = 1.314, df = 13, p = 0.106$). The `affixes + anchors` simulation correlates more strongly with phonological typicality than both the `anchors` simulation (Hotelling's t test: $t = 5.65, df = 13, p < 0.001$) and the `affixes` simulation (Hotelling's t test: $t = 4.071, df = 13, p < 0.001$). Finally, the `all` simulation has a stronger correlation between LDA log-odds and phonological typicality than the `neighbors + affixes` simulation (Hotelling's t test: $t = 2.335, df = 13, p < 0.05$) and than the `neighbors + anchors` simulation (Hotelling's t-test: $t = 4.696, df = 13, p < 0.001$). However, the correlation between LDA log-odds and phonological typicality observed in the `all` simulation is statistically indistinguishable from the same correlation observed in the `affixes + anchors` simulation (Hotelling's t test: $t = 0.38, df = 13, p = 0.355$). The results suggest that while the additional information about morpho-syntactic functions and anchor words improves correlations, the additional information about lexical categories however does not.

Results are mixed: when moving from simulations relying on a single source of information to the combination of two, we observe that all sources of information tend to be beneficial, increasing the correlation strength as compared to that achieved in simulations with a single source of information. The only exception is the addition of information about correlations with anchor words to the model which has information

about neighbors' lexical categories. In this case, the more complex model is not statistically different from the simpler model. However, when moving from models relying on two sources of information to the global model, we see that the information concerning the lexical categories of semantic neighbors is the only one which does not reliably increase the correlation between LDA log-odds and phonological typicality. This suggests that, once the simulation has access to the relation between estimated semantic vectors and real semantic vectors pertaining to both anchor words and morpho-syntactic functions, it can capture phonological typicality entirely. This is not the case when correlations with semantic vectors are combined with lexical category distributions in semantic neighborhoods. Even clearer is the preponderance of the morpho-syntactic semantic vectors, which have the strongest correlation by themselves and always improve a model when they are added to it. It seems to be the case that the relation between the semantics inferred from a non-word and that of highly grammaticalized morphological functions provides the most reliable source of information.

We can get further information by considering how LDA log-odds correlate with the behavioral results observed for children. Here, we see that correlations are lower, and differences across different simulations are weaker (Table 2). Again, all correlations are statistically reliable (except for the one concerning the `anchors` model[10]), suggesting that the information captured by the relation between non-words' semantics and the semantics encoded in the lexicon largely explains behavior. Once again, the strongest correlation is observed for the `affixes` simulation, although this correlation is statistically indistinguishable from the others. We repeated the comparisons across simulations, without detecting any statistically significant difference, hence no analogue of Table 3 is provided. The same comparisons were carried out, but all differences were not statistically reliable. Even with such a limited sample size, however, we find reliable

[10] Close inspection revealed an outlier in the distribution of the log-odds: a noun-like non-word has a much lower log-odds than the others, skewing results. We repeated the analyses removing this data point, and we observed a higher correlation, which was significant. However, this instability does not allow us to draw firm conclusions about the reliability of the correlation at hand.

indications that the way non-words interact with the lexicon from a semantic point of view largely accounts for the referential choices of children. Further experiments considering more non-words are expected to elucidate whether differences between different sources of information can be observed.

Finally, we focus on determining which individual dimensions contributed the most information to the clustering task. To this end, we used random forests, as previously described, relying on the default parameters of the *randomForest* R function. In order to evaluate the importance of a dimension, we considered the mean decrease in the Gini purity, which captures the degree to which splitting instances on a certain dimension leads to purer clusters. Results are summarized in Figure 2. For `neighbors`, nouns and verbs are the most informative, followed by prepositions and adjectives. Turning to `affixes`, we see that PLURAL and SINGULAR are much more informative than all other functions. For `anchors`, *dog, flower, carrot* and *sing* appear to be the most informative. It is interesting to note the prevalence of nouns, even though their corpus frequency is far lower.

When different sources of information are combined, we see a consistent picture. Verbs are again found to be the most informative, PLURAL and SINGULAR are the most useful inflections, *nose* and *sing* are the most useful anchor words. Interestingly, when affixes are combined with lexical categories (`neighbors + affixes`), the two most useful morpho-syntactic functions overshadow all lexical categories. This happens also when combining lexical categories with anchor words (`neighbors + anchors`), although the frequency distribution of verb neighbors is among the most useful dimensions. In `affixes + anchors`, morpho-syntactic functions and anchor words share the burden evenly, with *nose* and *sing* next to the usual PLURAL and SINGULAR. When all sources of information are combined, we observe the usual affixes, the verb category, and the verb *sing*. Taken together, we can infer from the results that the relations between the estimated semantic vectors and the semantics encoded in the lexicon are stable across simulations, which adds to their reliability. However, the informativity of specific dimensions likely depends on the target non-words: using

different stimuli may well highlight different dimensions in the semantic space.

Discussion

In this study we explored the hypothesis that there exists an (at least partially) systematic relation between word forms and their meanings, such that children can infer the gist of a word's semantics just from how it sounds. To evaluate this approach and the underlying hypothesis, we took non-words from the study by Fitneva et al. (2009), who investigated phonological bootstrapping Morgan and Demuth (1996) in the context of word learning. The goal was to explore whether the distinction based on phonological typicality (Farmer et al., 2006) between noun-like and verb-like non-words was also found in the semantic domain. If the semantic counterparts of phonologically distinctive non-words can be accurately clustered, the function mapping form onto meaning should have preserved similarity relations in the two domains. If, on the contrary, non-words cannot be reliably categorized on semantic grounds, it would follow that the form-to-meaning mapping orthogonalizes phonological and semantic vectors, such that similarity in one domain does not predict similarity in the other.

In detail, we generated an estimated semantic vector for each non-word using the LDL framework. Estimated semantic vectors were then correlated with the semantic vectors of real words, learned from a corpus of phonologically transcribed child-directed speech. In seven different simulations, Linear Discriminant Analysis (LDA) was used to cluster non-words into noun- or verb-like based on (i) the distribution of lexical categories of their semantic neighbors, (ii) their semantic relation to morpho-syntactic functions, (iii) their semantic relation with early acquired words, and all possible combinations of these three sources of information. Neighbors' lexical categories were observed to be the least informative source of information, while morpho-syntactic functions and anchor words were found to be mutually helpful, as measured by considering clustering accuracy, the discriminability introduced by semantic information, the correlation between the LDA log-odds and non-words' phonological typicality on the one hand, and behavioral patterns in the entity/action discrimination

task on the other.

The fact that neighbors' lexical categories provide the least information confirms our hypothesis that there is no need to explicitly operationalize an abstract lexical category to explain the results. On the contrary, lexical categories appear to be implicit in the relation between phonological forms and the semantic vectors they refer to. This conclusion follows from the following results. First, when neighbors' lexical categories were combined with either morpho-syntactic functions or anchor words a significant improvement in the correlation between LDA log-odds and phonological typicality was observed with respect to the simulations considering lexical categories alone, but not with respect to the simulations relying on either morpho-syntactic functions or anchor words alone. Moreover, the discriminability of the space was only statistically reliable when anchor words and morpho-syntactic functions were combined. As we already discussed, the low informativity of neighbors' lexical categories can be explained on the basis that by reducing each semantic neighbor to its lexical category, a lot of information is lost, since different semantic vectors are mapped to a same tag.

Interestingly, a strong correlation between phonological typicality scores and the confidence with which the LDA classified a non-word as a noun or a verb was found in all simulations. This suggests that phonology does not simply correlate with the lexical category of a word (Farmer et al., 2006; Fitneva et al., 2009) but also with its implicit meaning (Nygaard et al., 2009).

We note here that we do not assume that non-words have precise meanings that listeners become aware of, and that they can interpret and make explicit to themselves and others. Instead, we assume that the semantic vectors represent implicit meanings. The reason that these meanings must remain implicit is due to the fact that estimated semantic vectors are situated in portions of the semantic space which do not discriminate any precise experience, but, on the contrary, blend several types of experiences. The resulting blends appear nonetheless to be semantically coherent at a more abstract level, such that, for example, they make it possible to consistently recognize a non-word as denoting an entity or an action.

It is also important to note that semantic neighbors, and hence semantic relations, are only partially driven by phonology: it is true that the closest neighbor of the non-word /ɹɪsp/ is *crisp* and that the nearest neighbor or /hæps/ is *happy*, both with a very small edit distance[11]. However, the list of nearest neighbors of the first non-word also includes words such as *biscuit*, *raisin*, *chutney* and *waffle*, indicating that this non-word interacts with the dimensions of food and sweetness, regardless of the degree to which it shares phonological material with a word. In a similar way, the list of nearest neighbors for /hæps/ also includes *easy*, *lucky*, and *funny*. Here, the relation between phonology, semantics, and lexical category becomes apparent: nearest neighbors tend to have a consistent morpho-phonological feature, the final -Y, and tend to be adverbs derived from adjectives, even though the target non-word does not end in -Y. Moreover, the loose relation between phonology and semantic neighborhoods is confirmed by the observation that the nearest neighbor of many target non-words does not share phonological material with the target. This is, for example, the case of non-words /mɛfs/, /pɔsp/, /lɔfs/, /dwɪg/, and /sɪg/ whose nearest neighbors are *preferences*, *more*, *raviolis*, *cans*, and *strangers* respectively. Correlations of this kind are not assumed to be consciously available to speakers, but to influence their understanding in subtle ways.

Across the simulations we conducted, the confidence of the LDA classification was also significantly correlated with children's performances in the behavioral experiment by Fitneva et al. (2009), suggesting that semantic information estimated for non-words and its relation to the lexicon captures something relevant about how children picked the entity or action picture when hearing a non-word. Nonetheless, given that these results were obtained with only 16 non-words, designed to maximize the informativeness

---

[11] Fitneva and colleagues checked that the target non-words did not elicit consistent mappings to known words, such that participants could choose a referent based on the analogy between a target non-word and a similar sounding known word. While they found that undergraduates could come up with a similar word most of the times, these words did not show a consistent lexical category, such that the action/entity referential choice could be explained on the basis that a certain target non-word was found to be phonologically similar to a noun, and hence participants chose the entity referent analogizing over that word.

of their phonological form with respect to their lexical category, it will be informative to replicate these simulations with more non-words. Importantly, however, while the target non-words were not a random sample in the study by Fitneva et al. (2009), since they maximized discriminability on the phonological typicality dimension, the same target non-words are effectively a random sample in this study, since we did not optimize the choice of target non-words to maximize discriminability in semantic space. In any case, the correlations between the LDA's confidence and children's behavior suggest that semantic intuitions triggered by phonological forms play a role in explaining children's referential choices, even without explicitly modeling lexical categories. It is noteworthy that, according to the LDL-based model of the mental lexicon that we have used to explore nonword semantics, even upon its first occurrence and with no available contextual information, a word encounters a highly structured system which exploits all possible similarities in form and meaning to optimize lexical discrimination (see Baayen et al. (2018) for a detailed discussion of how such mappings explain inflectional systems).

An interesting issue that arises from this study concerns the *meaning of non-words*. With the LDL framework, the derivation of non-words' meaning becomes possible[12], and the estimated meanings generated by the model are far from random. Another study that implemented LDL to generate the estimated semantic vectors of non-words is Chuang et al. (submitted). By examining the acoustic durations and the response times of about 10 thousand auditory non-words in the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2018), they showed that the semantics of non-words influences the processing of auditory non-words to a substantial extent. Taken together, both the present study and Chuang et al. (submitted)

---

[12] See also the work by Hendrix presented at the 11^{th} *International Conference on the Mental Lexicon* in which *FastText* (Bojanowski, Grave, Joulin, & Mikolov, 2016) is used to generate semantic vectors for non-words and semantic neighborhood density is found to reliably predict reaction times in a lexical decision task. Non-words with denser semantic neighborhoods are found to be rejected more slowly, providing further evidence that non-words make contact with the lexicon. *FastText*, however, bears little relevance to cognitive modeling due to its built-in assumptions and its architecture.

demonstrate that non-words do make contact with the lexicon. That is, they are not semantically empty.

In a great number of behavioral experiments, non-words are commonly used to serve as controls or to avoid unwanted lexical and semantic effects. However, our results hint to the fact that non-words may elicit non-random semantic intuitions, which could possibly influence experimental results in a variety of tasks. Moreover, the idea that learners form an idea of what a word means from its form makes intuitive sense from an acquisition point of view (Nielsen & Rendall, 2014): a (phonotactically legitimate) non-word is simply a word a learner has not encountered yet. Many real words are non-words to somebody; actually, all words have been non-words to every learner during development. The meaning change of a non-word gradually becoming a word during the learning process is an interesting topic that merits further pursuit. This study also makes a testable prediction in the context of language acquisition and word processing. Since systematicity makes learning easier but hampers processing while arbitrariness enhances processing but impairs learnability (Blevins, Milin, & Ramscar, 2017b; Dautriche et al., 2017), it is predicted that the systematic form-to-meaning relation highlighted in the current study should be stronger for early acquired words and decrease as the vocabulary becomes larger (Gasser, 2004). This prediction is in line with the observation that sound-symbolic mappings facilitate word learning (Imai et al., 2008, 2015; Lockwood et al., 2016; Monaghan et al., 2012). Moreover, evidence from Monaghan (2014) showing that earlier acquired words are less subject to language change also strengthens our conclusions and the following prediction. If early acquired words show a more systematic form-to-meaning mapping, which is in turn beneficial to scaffold word learning, it makes sense that those words are less subject to language change, since they encode the sound-to-meaning correspondences which are necessary to language learning. In other words, if children rely on a systematic sound-to-meaning mapping to bootstrap word learning, we would precisely expect those words which first encode systematic mappings to be somehow protected. The fact that Monaghan (2014) reports precisely this pattern strengthens our argument. Summing up, our results hint

to the possibility that systematic form-to-meaning relations extend beyond sound-symbolism (Dingemanse et al., 2015) and also affect the structuring of the mental lexicon along dimensions which can give rise to categorically constrained behavior even in the absence of abstract categories.

At a very speculative level, word learning may happen in the following way. On the one hand, an estimated semantic vector is derived from the sound input. At this point, syntactic, morphological, and lexical biases — such as those documented in the reported simulations — immediately come into play, together with other factors such as valence, arousal, and dominance (Westbury, 2014). On the other hand, and at the same time, there is the cognitive understanding of the world that needs to be taken into consideration. At this point, two options are possible. One is to consider current text-based semantic vectors as proxies for more rich vectors which integrate linguistic and perceptual information (Bruni, Tran, & Baroni, 2014). Under this view, the real world input would have already co-determined the location of the estimated semantic vector in the semantic space. Alternatively, linguistic and perceptual semantic vectors are kept distinct and they differentially affect the estimated semantic vector. However, this second view requires further specification of how exactly linguistic and real world information are linked. We favor the first option, since it is simpler. Moreover, under this view, the semantic vectors we have used are necessarily incomplete as they are derived from purely linguistic data. Yet, in spite of this imperfection, results are robust.

In conclusion, our study targets a further possible pocket of systematicity in the relation between form and meaning, next to onomatopoeia, sound symbolism (Hinton et al., 1994; Imai & Kita, 2014; Maurer et al., 2006; Nuckolls, 1999; Sapir, 1929; Sidhu & Pexman, 2015; Westbury et al., 2017), and phonaestemes (Bergen, 2004; Pastizzo & Feldman, 2009; Åsa, 1999), which concerns the relation between phonology and the likely referent of a word. Our work builds on two lines of research. On the one hand, the phonological bootstrapping literature, which highlights robust and informative correlations between the phonology of a word and its lexical category (Morgan & Demuth, 1996). However, contrary to this account and in line with studies on the

non-arbitrary nature of the form-to-meaning relation (Dingemanse et al., 2015; Monaghan et al., 2014; Sidhu & Pexman, 2017), we show that a similarly informative relation may exist between phonology and meaning, such that it is not necessary to postulate the existence of abstract lexical categories to explain the referential choices of the children who participated in the experiment by Fitneva et al. (2009). We tested this possibility using non-words and showed that once a non-word is projected into semantic space, this semantic space captures the same distinctions as the phonological domain, suggesting a surprisingly systematic mapping between the two. Lexical categories appear to be captured by discrimination-driven semantic vectors and a straightforward linear mapping between these vectors and the phonological form of the corresponding words. Our results are in line with those of other studies on the semantics of non-words (Chuang et al., submitted), calling for further research to evaluate the hypothesis that phonological patterns do indeed influence our semantic percept of novel words.

The evolutionary advantage of this partially predictable form-to-meaning mapping stems from the following consideration. While being beneficial for processing, since it maximises discriminability of forms (Dingemanse et al., 2015; Gasser, 2004), a completely arbitrary mapping makes learning costly and inefficient, since generalizations are never warranted and every mapping has to be learned individually (Nygaard et al., 2009). Therefore, it is possible that languages evolve in such a way to encode superficial, coarse meaning features in the sound patterns of a word Adelman et al. (2018); Chater and Christiansen (2010); Monaghan, Christiansen, Farmer, and Fitneva (2011); Monaghan et al. (2012, 2014), such that it is possible to infer a first gist of a word's meaning from the presentation of the word's form. It follows that new words, real or nonse, are never orthogonal to the lexical system, and should not be conceptualized as entries that are added into a list, into a slot that effectively is a tabula rasa. Instead, any new form generates, albeit implicitly, expectations about its semantics.

Acknowledgments

The first author was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp; the second and third authors were supported by a European Research Council grant (#742545 (WIDE)) awarded to the third author. We thank Konstantin Sering and Fabian Tomaschek for useful discussion, Peter Hendrix and two anonymous reviewers for their insightful comments and suggestions.

Appendix

As shown in Table 4, the simulation considering frequency counts over semantic neighbors' lexical categories did not present any collinearity issue, hence the full input space was used. For the simulation considering morpho-syntactic functions, 12 affixes were preserved: eight are inflectional (COMPARATIVE, SUPERLATIVE, CONTINUOUS, PERSISTENCE, PRESENT, PAST, SINGULAR, and PLURAL), two are derivational (-FUL and -Y), and two are modal functions (CAN and SHALL). As far as the simulation considering anchor words is concerned, 13 anchor words were retained: seven verbs (*cry, eat, see, try, get, sing, sit*) and six nouns (*daddy, carrot, dog, ball, nose, flower*).

When combining together the sources of information, the adverse effect of collinearity increased. Keep in mind that the threshold is defined for the variance of z-scores while the previous thresholds addressed the variance of correlation vectors, hence the different order of magnitude. For all simulations, 14 dimensions were retained after setting the threshold. Within the simulation combining neighbors' lexical categories and morpho-syntactic functions, 3 lexical categories (verb, noun, and adjective) and 11 affixes (CAN, COMPARATIVE, CONTINUOUS, -FUL, PAST, PLURAL, PRESENT, SINGULAR, SHALL, SUPERLATIVE, and -Y) were retained. When the lexical categories were combined with the anchor words, the same 3 lexical categories as in the previous simulation and 11 anchor words *daddy, dog, ball, nose, flower, cry, see, try, get, sing, sit* survived the cutoff. The combination of morpho-syntactic functions and anchor words made use of 5 anchor words (three verbs, *cry, get, sing*, and two nouns *daddy, nose*) and 9 affixes (COMPARATIVE, CONTINUOUS, -FUL, PAST, PLURAL,

SINGULAR, SHALL, SUPERLATIVE, and -Y). Finally, when all sources of information were combined, 4 anchor words (*daddy, nose, get, sing*), 2 lexical categories (*verbs, adjectives*), and 8 morpho-syntactic functions (COMPARATIVE, -FUL, PAST, PLURAL, SINGULAR, SHALL, SUPERLATIVE, -Y) survived the cutoff.

References

Adelman, J. S., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes [Journal Article]. *Cognition*, *175*, 122-130. doi: 10.1016/j.cognition.2018.02.007

Ambridge, B. (2017). Syntactic categories in child language acquisition: Innate, induced, or illusory? [Book Section]. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., p. 567 - 580). Elsevier.

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings [Journal Article]. *The mental lexicon*, *13*(2), 230-268. doi: 10.1075/ml.18010.baa

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). Discriminative morphology: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in discriminative learning [Journal Article]. *Complexity*, *2019*, 1 - 39. doi: 10.1155/2019/4895891

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*. doi: 10.1080/02687038.2016.1147767

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

Blevins, J. P., Milin, P., & Ramscar, M. (2017a). The zipfian paradigm cell filling problem. *Perspectives on Morphological Organization: Data and Analyses*, *10*, 141.

Blevins, J. P., Milin, P., & Ramscar, M. (2017b). The zipfian paradigm cell filling problem. *Perspectives on morphological organization: Data and analyses*, *10*, 141.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics [Journal Article]. *Journal of artificial intelligence research*, *49*(1), 1-47. doi: 10.1613/jair.4135

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments [Journal Article]. *Journal of memory and language*, *30*(3), 348-369. doi: 10.1016/0749-596x(91)90041-H

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution [Journal Article]. *Cognitive science*, *34*(7), 1131-1157. doi: 10.1111/j.1551-6709.2009.01049.x

Christophe, A., Guasti, T., Nespor, M., Dupoux, E., & Van Ooyen, B. (1997). Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition [Journal Article]. *Language and cognitive processes*, *12*(5-6), 585-612. doi: Doi 10.1080/016909697386637

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition [Journal Article]. *Language and speech*, *51*(1-2), 61-75. doi: 10.1177/00238309080510010501

Chuang, Y.-Y., Voller, M.-I., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (submitted). The processing of nonword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning [Journal Article].

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing* [Book]. Cambridge, UK: Cambridge University Press. doi: 10.2277/0521808901

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities [Journal Article]. *Cognition*, *163*, 128-145. doi: 10.1016/j.cognition.2017.02.001

de Carvalho, A., He, A. X., Lidz, J., & Christophe, A. (2019). Prosody and function words cue the acquisition of word meanings in 18-month-old infants [Journal Article]. *Psychological Science*, 956797618814131. doi: 10.1177/0956797618814131

de Saussure, F. (1916). *Course in general linguistics* [Book]. New York, NY: McGraw-Hill.

Diedenhofen, B., & Musch, J. (2015). cocor: a comprehensive solution for the statistical

comparison of correlations [Journal Article]. *PLoS One*, *10*(3), e0121945. doi: 10.1371/journal.pone.0121945

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. doi: 10.1016/j.tics.2015.07.013

D'Onofrio, A. (2013). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm [Journal Article]. *Language and Speech*, *57*(3), 367-393. doi: 10.1177/0023830913507694

Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test [Book Section]. In B. Höhle & J. Weissenborn (Eds.), *Approaches to bootstrapping: Phonological, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, p. 189 - 232). Amsterdam, The Netherlands: Benjamin.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension [Journal Article]. *Proceedings of the National Academy of Sciences*, *103*(32), 12203-8. doi: 10.1073/pnas.0602173103

Fenson, L., Marchman, V. A., Thal, D., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *Macarthur-bates communicative development inventories: User's guide and technical manual* (2nd ed.) [Book]. Baltimore, MD: Brookes Publishing Co.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951* [Book]. Oxford, UK: Oxford University Press.

Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: phonological constraints on children's lexical categorization of new words [Journal Article]. *Journal of child language*, *36*(5), 967-97. doi: 10.1017/S0305000908009252

Gasser, M. (2004). The origins of arbitrariness in language [Book Section]. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Gillis, S., & Ravid, G. (2009). Language acquisition [Book Section]. In D. Sandra,

J.-O. Östman, & J. Verschueren (Eds.), *Cognition and pragmatics* (p. 201-249). Amsterdam: Benjamin.

Gleitman, L. R. (1990). The structural sources of verb meanings [Journal Article]. *Language acquisition*, *1*(1), 3-55. doi: 10.1207/s15327817la0101_2

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork [Journal Article]. *Cognition*, *52*(2), 125-157. doi: 10.1016/0010-0277(94)90065-5

Harrell Jr, F. E., with contributions from Charles Dupont, & many others. (2017). Hmisc: Harrell miscellaneous [Computer software manual]. (R package version 4.0-3)

Hinton, L., Nichols, J., & Ohala, J. J. (1994). *Sound symbolism* [Book]. Cambridge, UK: Cambridge University Press.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution [Journal Article]. *Philosophical transactions of the Royal Society London B - Biological sciences*, *369*(1651). doi: 10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65.

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015, 02). Sound symbolism facilitates word learning in 14-month-olds. *PLoS ONE*, *10*(2), 1-17. doi: 10.1371/journal.pone.0116494

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in english-speaking children. *Cognitive science*, *35*(3), 575-586. doi: 10.1111/j.1551-6709.2010.01169.x

Kelly, M. H. (1988). Phonological biases in grammatical category shifts [Journal Article]. *Journal of memory and language*, *27*(4), 343-358. doi: 10.1016/0749-596x(88)90060-5

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand english words [Journal Article]. *Behavior research*

methods, *44*(4), 978-90. doi: 10.3758/s13428-012-0210-4

Köhler, W. (1929). *Gestalt psychology* [Book]. New York, NY: Liveright.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge [Journal Article]. *Psychological review*, *104*(2), 211-240. Retrieved from `<Go to ISI>://WOS:A1997WU96300001` doi: 10.1037/0033-295X.104.2.211

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lockwood, G., Dingemanse, M., & Hagoort, P. (2016). Sound-symbolism boosts novel word learning [Journal Article]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1274-1281. doi: 10.1037/xlm0000235

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence [Journal Article]. *Behavior research methods, instruments & computers*, *28*(2), 203-208. doi: 10.3758/Bf03204766

MacWhinney, B. J. (2000). *The childes project: Tools for analyzing talk. the database.* (3rd ed., Vol. 2) [Book]. Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children syntax: The nature and ontogenesis of syntactic categories [Book Section]. In K. E. Nelson (Ed.), *Children's language* (Vol. 2, p. 127-213). New York, NY: Gardner Press.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults [Journal Article]. *Developmental science*, *9*(3), 316-22. doi: 10.1111/j.1467-7687.2006.00495.x

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Computer software manual]. (R package version 1.6-8)

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017, feb). Discrimination in lexical decision. *PLoS ONE*, *12*(2), e0171935. doi: 10.1371/journal.pone.0171935

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution [Journal

Article]. *Cognition*, *133*(1), 530-534. doi: 10.1016/j.cognition.2014.08.0079

Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2011). Measures of phonological typicality: Robust coherence and psychological validity [Journal Article]. *The mental lexicon*, *5*(3), 281-299. doi: 10.1075/ml.5.3.02mon

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: learning advantages from the structure of the vocabulary [Journal Article]. *Journal of experimental psychology general*, *140*(3), 325-47. doi: 10.1037/a0022924

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning [Journal Article]. *Journal of experimental psychology: Learning, memory, and cognition*, *38*(5), 1152-64. doi: 10.1037/a0027747

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? [Journal Article]. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1651), 20130299. doi: 10.1098/rstb.2013.0299

Morgan, J. L., & Demuth, K. (1996). Signal to syntax: An overview [Book Section]. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (p. 1-24). Mahwah, NJ: Erlbaum.

Nielsen, A., & Rendall, D. (2014). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission [Journal Article]. *Language and cognition*, *4*(02), 115-125. doi: 10.1515/langcog-2012-0007

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification [Journal Article]. *Journal of Mathematical Psychology*, *34*(4), 393-418. doi: 10.1016/0022-2496(90)90020-A

Nuckolls, J. B. (1999). The case for sound-symbolism [Journal Article]. *Annual review of anthropology*, *28*, 225 - 252. doi: 10.1146/annurev.anthro.28.1.225

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning [Journal Article]. *Cognition*, *112*(1), 181-6. doi:

10.1016/j.cognition.2009.04.001

Pastizzo, M. J., & Feldman, L. B. (2009). Multiple dimensions of relatedness among words: Conjoint effects of form and meaning in word recognition [Journal Article]. *The mental lexicon*, *4*(1), 1. doi: 10.1075/ml.4.1.01pas

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

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia - a window into perception, thought and language [Journal Article]. *Journal of Consciousness Studies*, *8*(12), 3-34.

Ramscar, M., & Port, R. (2015). Categorization (without categories) [Book Section]. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (p. 75-99). Berlin/Boston: De Gruyter Mouton.

Sapir, E. (1929). A study in phonetic symbolism [Journal Article]. *Journal of experimental psychology*, *12*(3), 225 - 239. doi: 10.1037/h0070931

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing.* Manchester, UK.

Sharpe, V., & Marantz, A. (2017). Revisiting form typicality of nouns and verbs [Journal Article]. *The Mental Lexicon*, *12*(2), 159-180. doi: 10.1075/ml.17004.sha

Sharpe, V., Reddigari, S., Pylkkänen, L., & Marantz, A. (2018). Automatic access to verb continuations on the lexical and categorical levels: Evidence from meg [Journal Article]. *Language, Cognition and Neuroscience*, *34*(2), 137 - 150. doi: 10.1080/23273798.2018.1531139

Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? sound symbolism and gender in first names [Journal Article]. *PLoS One*, *10*(5), e0126809. doi: 10.1371/journal.pone.0126809

Sidhu, D. M., & Pexman, P. M. (2017, Oct 01). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, *25*(5), 1619 - 1643. doi: 10.3758/s13423-017-1361-1

Sloutsky, V. M. (2003). The role of similarity in the development of categorization [Journal Article]. *Trends in cognitive sciences*, *7*(6), 246-251. doi: 10.1016/S1364-6613(03)00109-8

Styles, S. J., & Gawne, L. (2017). When does maluma/takete fail? two key failures and a meta-analysis suggest that phonology and phonotactics matter [Journal Article]. *Iperception*, *8*(4), 2041669517724807. doi: 10.1177/2041669517724807

Todorov, V., & Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis [Journal Article]. *Journal of Statistical Software*, *32*(3). doi: 10.18637/jss.v032.i03

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The massive auditory lexical decision (mald) database [Journal Article]. *Behavior research methods*. doi: 10.3758/s13428-018-1056-1

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. (ISBN 0-387-95457-0)

Westbury, C. (2014). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of psycholinguistic research*, *43*(5), 631–649.

Westbury, C., & Hollis, G. (2018). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging [Journal Article]. *Behavioral Research Methods*. doi: 10.3758/s13428-018-1118-4

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2017). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength [Journal Article]. *Journal of memory and language*, *99*, 122-150. doi: 10.1016/j.jml.2017.09.006

Åsa, A. (1999). *Studies in sound symbolism* (Thesis). Göteborg University.

Source	Acc		Errors			Wilks' λ			
		p value		C	P		df	χ^2	p value
neighbors	15/16	< 0.001	sig	V	N	0.409	9	8.505	0.484
affixes	14/16	< 0.01	risp	N	V	0.281	12	10.156	0.602
			skik	V	N				
anchors	16/16	< 0.001				0.335	13	8.194	0.831
neighbors + affixes	16/16	<0.001				0.045	14	21.661	0.086
neighbors + anchors	16/16	< 0.001				0.245	14	9.847	0.773
anchors + affixes	16/16	< 0.001				0.004	14	38.246	< 0.001
all	16/16	< 0.001				0.005	14	36.684	< 0.001

Table 1

Categorization and discrimination statistics for the seven simulations. Source indicates the source of information or combinations thereof that were used as input to the LDA. `Neighbors` *indicates the lexical categories of the semantic neighbors,* `affixes` *indicates the morpho-syntactic functions,* `anchors` *indicates the anchor words, and* `all` *indicates the combination of all three sources. Acc: accuracy in the clustering task expressed as the ratio of correctly clustered non-words to the total number of non-words. The corresponding p-value is computed using an exact binomial test considering chance as the null hypothesis. The column C indicates the intended lexical category of a non-word based on its phonology. The column P indicates the lexical category predicted for a non-word by the LDA. The column df indicates the degrees of freedom of the corresponding χ^2 distribution. The column χ^2 indicates the value of the χ^2 statistics corresponding to Wilks' λ.*

Source	LDA log-odds ∼Typicality		LDA log-odds ∼p(Action)	
		p-value		p-value
neighbors	0.734	< 0.001	0.529	< 0.05
affixes	0.871	< 0.001	0.603	< 0.05
anchors	0.800	< 0.001	0.481	0.06
neighbors + affixes	0.961	< 0.001	0.658	< 0.01
neighbors + anchors	0.861	< 0.001	0.511	< 0.05
affixes + anchors	0.981	< 0.001	0.655	< 0.01
all	0.984	< 0.001	0.665	< 0.01

Table 2

Correlation statistics for the seven simulations. Source indicates the source of information or combinations thereof that were used as input to the LDA. `Neighbors` *indicate the lexical categories of the semantic neighbors,* `affixes` *indicate the morpho-syntactic functions,* `anchors` *indicate the anchor words,* `all` *indicate the combination of all three sources. LDA log-odds ∼Typicality indicates the correlation between the confidence of the LDA in clustering a target non-word and the degree of phonological typicality of the non-word itself. LDA log-odds ∼p(Action) indicates the correlation between the LDA log-odds and the proportion of children who selected the action referent when presented with a target non-word.*

Simulation 1	Simulation 2	corr 1	corr 2	Hotelling T	df	p
neighbors+ affixes	neighbors affixes	0.961	0.734	4.248	13	<0.001
neighbors+ anchors	neighbors anchors	0.861	0.734	1.314	13	0.106
		0.861	0.800	2.603	13	<0.05
anchors+ affixes	anchors affixes	0.981	0.800	5.650	13	<0.001
		0.981	0.871	4.071	13	<0.001
all	neighbors+ affixes	0.984	0.961	2.335	13	<0.05
	neighbors+ anchors	0.984	0.861	4.696	13	<0.001
	anchors+ affixes	0.984	0.981	0.380	13	0.355

Table 3

Comparisons of the correlation between phonological typicality and LDA log-odds across pairs of simulations. The first column lists the more complex model. The second column lists the less complex model. The third and fourth columns show the correlation between LDA log-odds and phonological typicality in the first and second simulation, respectively. The two correlations in a same row are compared to test whether the correlation obtained in Simulation 1 is significantly larger than the correlation obtained in Simulation 2. The Hotelling T statistic and relative p-value, thus, tell whether the null hypothesis that the more complex model is worse than or indistinguishable from the more simple model can be rejected.

	Original dimensions	Threshold	Remaining Dimensions
neighbors	9	-	9
affixes	43	0.0020	12
anchors	20	0.0022	13
neighbors + affixes	51	0.7	14
neighbors + anchors	29	0.55	14
affixes + anchors	63	0.99	14
all	81	1.25	14

Table 4

Number of input dimensions for the LDA analyses across different simulations. The first column provides the dimensionality of the original input space. The second column indicates the threshold on the variance of input dimensions chosen to prevent collinearity issues in the LDA. The third column indicates the maximum number of remaining non-collinear dimensions. The threshold for the first three simulations is defined on the variance of the original dimensions, while the threshold for the four last simulations is defined on the z-transformed input dimensions, hence the different order of magnitude.

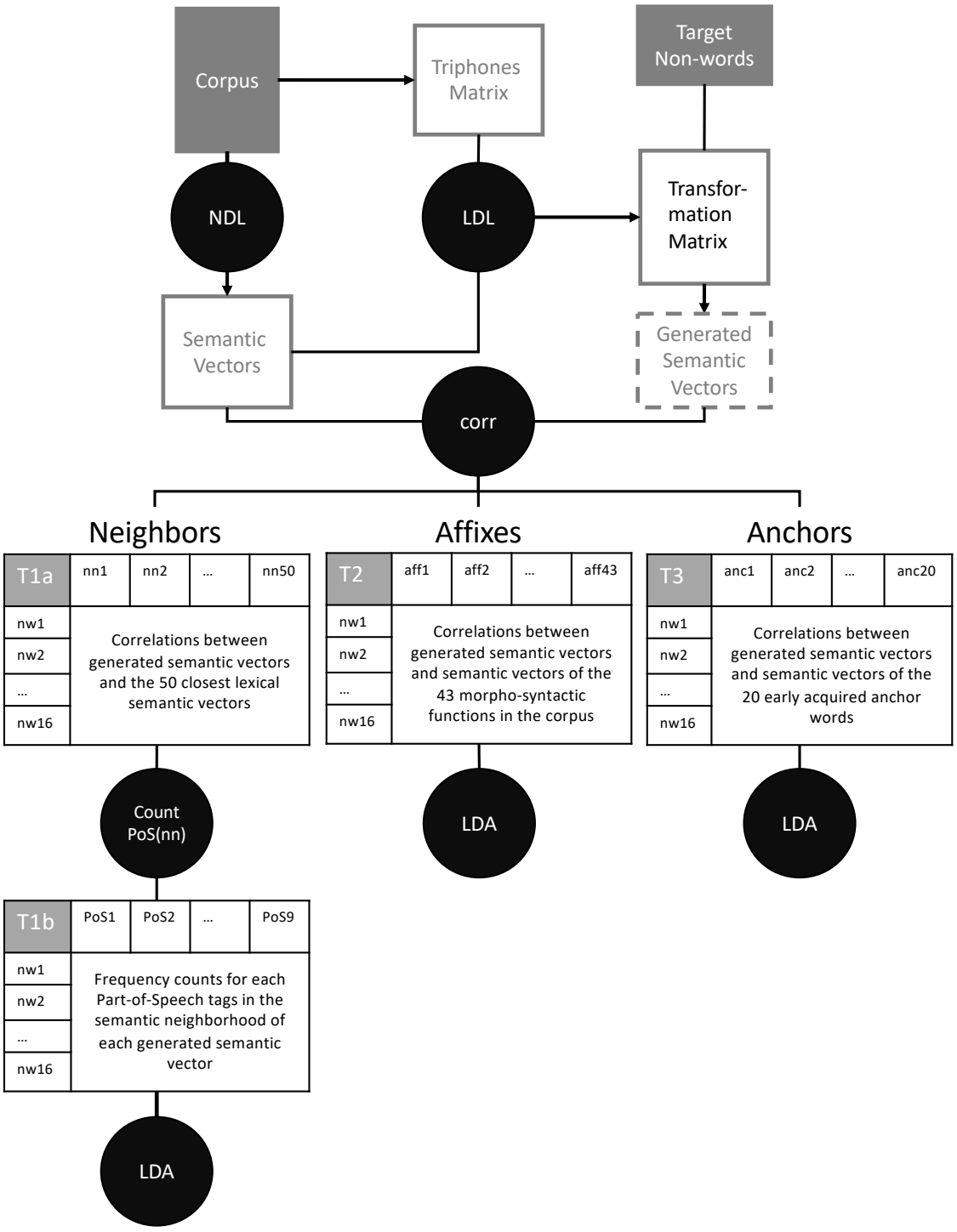

Figure 1. Schematic depiction of the experimental design. Elements in gray indicate resources, black circles indicate algorithms and processes, white boxes indicate data structures generated through the process and used for the experiment.

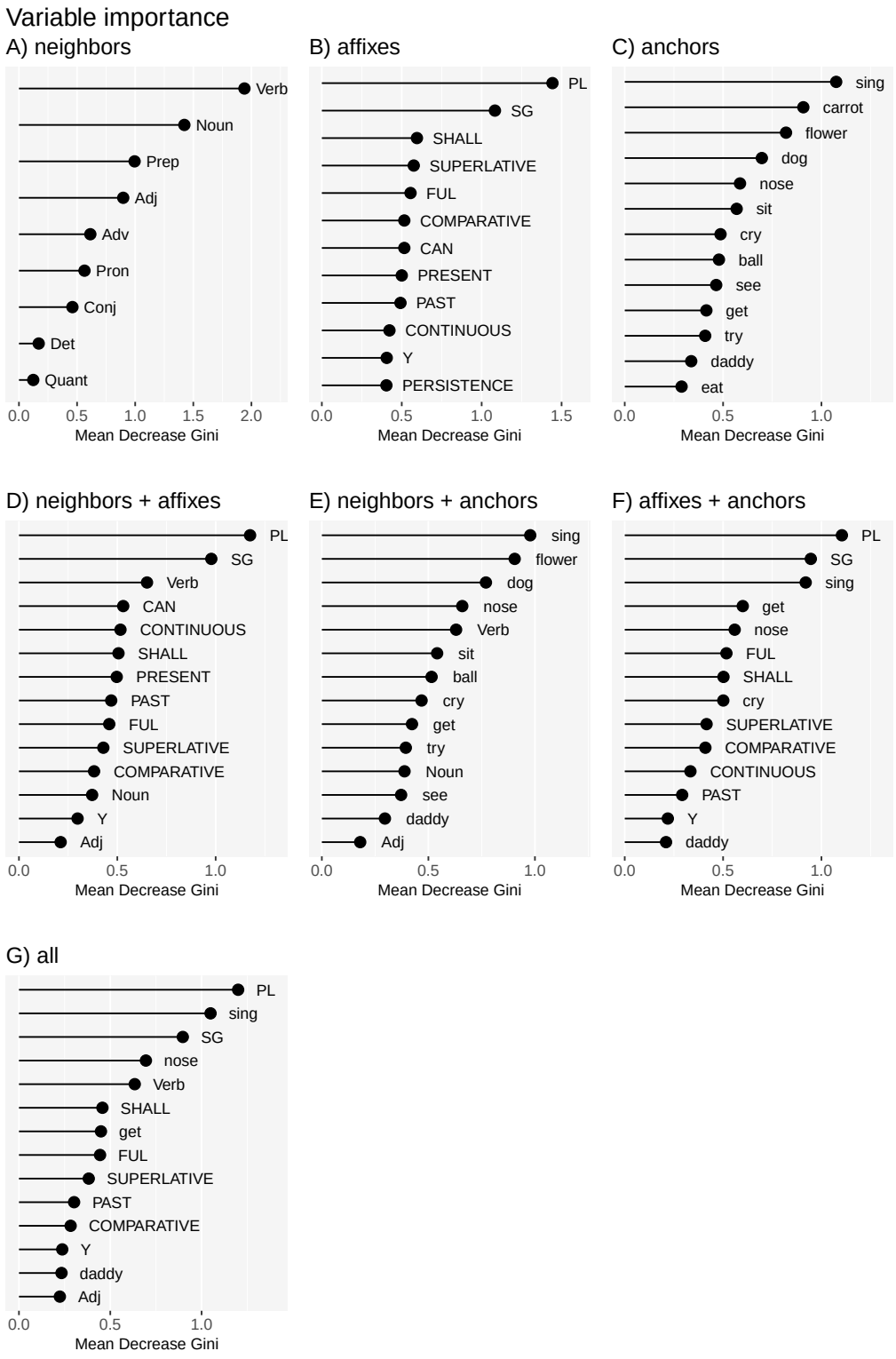

Figure 2. Variable importance as measured using the Mean Decrease in Gini purity, obtained with the use of Random Forests. The top line shows the simulations relying on 1 source of information. The mid-row shows the simulations relying on 2 sources of information. The last plot shows the simulation which relies on all sources of information.