Division of Labor in Vocabulary Structure: Insights from Corpus Analyses

Morten H. Christiansen[*]

Department of Psychology, Cornell University, Ithaca, NY, USA

Haskins Laboratories, New Haven, CT, USA

Padraic Monaghan

Centre for Research in Human Development and Learning, Department of Psychology,

Lancaster University, Lancaster, UK

Keywords: Vocabulary; corpus analysis; form-meaning mappings; arbitrariness; systematicity; lexical categories; sound symbolism; big data

Total word count: 7,021

[*]Please address correspondence to:
Morten H. Christiansen
Department of Psychology
228 Uris Hall
Cornell University
Ithaca, NY 14853
Phone: 607-255-3834
e-mail: christiansen@cornell.edu

Abstract

Psychologists have used experimental methods to study language for more than a century. However, only with the recent availability of large-scale linguistic databases has a more complete picture begun to emerge of how language is actually used and what information is available as input to language acquisition. Analyses of such ‘big data’ have resulted in reappraisals of key assumptions about the nature of language. As an example, we focus on corpus-based research that has shed new light on the arbitrariness of the sign: the longstanding assumption that the relationship between the sound of a word and its meaning is arbitrary. The results reveal a systematic relationship between the sound of a word and its meaning, which is pronounced for early acquired words. Moreover, the analyses further uncover a systematic relationship between words and their lexical categories—nouns and verbs sound differently from each other—affecting how we learn new words and use them in sentences. Together, these results point to a division of labor between arbitrariness and systematicity in sound-meaning mappings. We conclude by arguing in favor of including ‘big data’ analyses into the language scientist’s methodological toolbox.

1. Introduction

Since Wilhelm Wundt established the first experimental psychology laboratory in 1879 at the University of Leipzig (Boring, 1960), psychologists have sought to understand the mind through a variety of empirical methods. Using experimental methods, psychologists have subsequently gained substantial insight into the unique human ability for language. However, just as studying gratings and black bars on a computer screen does not reveal the full complexity of the visual world and how our brains might be dealing with it (Olshausen & Field, 2005), so do the traditional lab-based psycholinguistic experiments not capture the full nature of language and its impact on the human mind. One way of getting a more comprehensive picture of language in the real world is to incorporate 'big language data' in the form of corpora into the psychology of language.

The study of language corpora has a long historical pedigree going back more than a century. When it comes to language use, corpora were employed initially primarily to determine various distributional properties of language (see McEnery & Wilson, 1996, for a review). For example, Kading (1897) used a corpus consisting of 11 million words to determine the frequency of letters and sequences thereof in German. Corpora were subsequently collected and used by field linguists (e.g., Boas, 1940) and structuralist linguists (e.g., Harris, 1951; Hockett, 1954) to inform their linguistic theories. However, this work was limited by their treatment of corpora simply as collections of utterances that could be subjected only to relatively simple bottom-up analyses. A more comprehensive approach to corpora was proposed by Quirk (1960) in his introduction of the still ongoing project, *The Survey of English Usage*. From the viewpoint of psychology, the compilation of the 1 million word *Brown Corpus* (Kucera & Francis, 1967) in the 1960s was a major milestone. In particular, the use of computers to analyze the corpus resulted in word frequency statistics that were used to control psycholinguistic studies until quite recently. Currently, word frequency is generally assessed using much larger corpora, such as the British National Corpus (Burnard & Aston, 1998), COCA (Davies, 2010), or the Google Terabyte Corpus (Brants & Franz, 2006; which provides a snapshot of the world-wide-web in 2006). Thus, use of corpus data for the purpose of stimulus control is now standard in psycholinguistic experimentation.

There is also a long history of using corpora in the study of language acquisition (for reviews, see Behrens, 2008; Ingram, 1989). Much of this history is characterized by diary studies of children, many of which tended to focus on development in general rather than on language per se. Indeed, even Charles Darwin (1877) wrote a paper on the development of his infant son. Modern diary studies have generally concentrated more on children's productions of specific aspects of language such as errors in argument structure (Bowerman, 1974) or verb use (Tomasello, 1992). Other studies have approached children's productions by collecting extended samples longitudinally from multiple children (e.g., Bloom, 1970; Braine, 1963). Brown's (1973) longitudinal study of three children—Adam, Eve, and Sarah—constitutes an important milestone by the use of tape recordings. Importantly, the transcriptions of the three children's language samples across their early linguistic development eventually became part of the *Child Language Data Exchange System* (CHILDES, MacWhinney, 2000), which was originally conceived by Catherine Snow and Brian MacWhinney in 1984 as a central depository of language acquisition corpus data (MacWhinney & Snow, 1985). The CHILDES database, as a source of (relatively) big data, has become the most prominent source of language acquisition data for both sophisticated statistical analyses and computational modeling. As such, corpus data has become a key component of developmental psycholinguistics.

Today, corpora constitute an integral part of psychological studies of language acquisition and use. Analyses of linguistic databases have lead to reappraisals of key assumptions about the nature of language (e.g., regarding what is available in the input to children). As an example, we focus on corpus-based research that has shed new light on the arbitrariness of the sign: the longstanding assumption that the relationship between the sound of a word and its meaning is arbitrary (for reviews of non-corpus work relating to onomatopoeia, ideophones, phonaesthemes, sign language iconicity, as well as sound-shape and sound-affect correspondences, see Perniss, Thompson & Vigliocco, 2010; Schmidtke, Conrad & Jacobs, 2014). In what follows, we first discuss results from analyses of English, indicating that language incorporates a statistically significant amount of systematicity in form-meaning correspondences across the vocabulary. We then consider further corpus analyses suggesting that additional systematicity can be found at the level of lexical categories, revealing the 'sound of syntax'. Results from human experimentation corroborate the corpus analyses, pointing to a division of labor between arbitrariness and systematicity in the structure of the vocabulary. We conclude that language scientists need to embrace 'big data' in order to get a full picture of how language works.

2. How arbitrary is spoken language?

Ever since Saussure (1916) famously noted that *"le signe est arbitraire"* (p. 100), it has been assumed that the relationship between the sound of a word and its meaning is arbitrary. Indeed, Hockett (1960) selected the arbitrariness of the sign as one of the defining features of human language. The assumption of form-meaning arbitrariness is fundamental to most modern grammatical theories on both sides of the Chomskyan divide. For example, Pinker (1999, p. 2) states that “onomatopoeia and sound symbolism certainly exist, but they are asterisks to the far more important principle of the arbitrary sign—or else we would understand the words in every foreign language instinctively, and never need a dictionary for our own!” In a similar vein, Goldberg (2006, p. 217) notes that “... the particular phonological forms that a language chooses to convey particular concepts [...] generally are truly arbitrary, except in relative rare cases of phonaesthemes.” When considering that the perennial woody plant that we refer to in English as *tree* is called *Baum* in German, *arbre* in French, and *shù* (樹) in Mandarin, the arbitrariness of the sign seems obvious. Even onomatopoeia can appear seemingly arbitrary: the sound that pigs make are called *oink oink* in English, *ut it* in Vietnamese, *kvik kvik* in Czech, and *øf øf* in Danish.

Historically, however, the idea of the arbitrariness of the sign has been far from obvious. In fact, throughout most of human intellectual history, from the Greek philosophers through to the Renaissance and Enlightenment scholars (for a review, see Eco, 1995), the sound of a word was often assumed to directly express its meaning. This is exemplified by the 2,300-year-old debate between Hermogenes and Cratylus over whether the nature of referents inheres within words (Hamilton & Cairns, 1961). Recent research on sound symbolism has revealed that at least some word forms have a systematic relationship to their meanings. This systematic relationship may appear either as *absolute iconicity*, where a linguistic feature directly imitates some aspect of semantics (as in onomatopoeia) or *relative iconicity*, where there may be statistical regularities between similar sounds and similar meanings in the absence of imitation (Gasser, Sethuraman & Hockema, 2011). An example of the latter type of systematic form-meaning mapping can be found in phonaesthemes; e.g., the tendency in English for words ending in *-ump* to refer to rounded things such as *lump*, *bump*, *mump*, and *rump* (something that can even be primed in native English speakers; Bergen, 2004). Thus, some non-arbitrariness does seem to exist in form-meaning mappings—but just how arbitrary is language?

Monaghan, Shillcock, Christiansen and Kirby (in press) employed corpus analyses to determine the degree of arbitrariness in language. They extracted phonological forms for all the English monosyllabic words found in the CELEX database (Baayen, Pipenbrock & Gulikers, 1995), accounting for about 70% of words used in English (Baayen et al. 1995). To ensure that potential systematicity in form-meaning mappings was not due to the specific ways in which either form or meaning similarity was represented, Monaghan et al. employed several different methods. They computed sound similarity between word forms in three different ways: a) *phonological feature edit distance*: the minimum number of phonological feature changes required to convert one word to another (e.g., *cat* and *dog* differ by 8 features, associated with manner and place of articulation); b) *phoneme edit distance*: the number of phoneme changes required to convert one word to another (e.g., *cat* and *dog* differ by 3 phonemes); and c) *phonological feature Euclidean distance*: the Euclidean distance between phonological feature representations of words (e.g., *cat* and *dog* differ by a distance of .881). Two different representations of meaning[1] were used to compute meaning similarity: a) *contextual co-occurrence vectors* generated by counting words appearing within a +/-3 word window with each of 446 context words in the British National Corpus (Burnard & Aston, 1998); and b) *semantic features derived from WordNet* (Miller, 2013), in which words are grouped together according to

[1] Both types of semantic representations have been used extensively in computational linguistics, in part reflecting behavioral responses to meaning similarity (e.g., Huettig, Quinlan, McDonald & Altmann, 2006). Contextual co-occurrence vectors capture the tendency for words with similar meanings to occur in similar contexts, thus resulting in similar vectors (see Riordan & Jones, 2011, for a review of different ways of computing such vectors). In contrast, WordNet aims to capture similarity between word meanings in terms of hyponymy, that is, words are defined in terms of so-called *is-a* relations; e.g., a *dog* is-a canine, which is-a carnivore, which is-a mammal, and so on (for an introduction, see Fellbaum, 2005).

hierarchical relations and grammatical properties (e.g., *cat* and *dog* share 13 features [*entity, organism, animal, vertebrate, mammal, placental, carnivore, has paws, has tail, has ribs, has thorax, has head, has face*] and differ by 8 features, including *feline* versus *canine*).

Using these representations, Monaghan et al. (in press) generated separate similarity spaces for sound and meaning by comparing the representation for a given word to all other words. For example, for the word *dog*, similarity measures would be computed between its phonological representation and the sound representations of all other words. Likewise, the meaning representation of *dog* would be compared with the semantic representations of all the other words in the meaning similarity space. This produces the same number of similarity pairs (such as *dog* – *cat*, *dog* – *cog*, etc.) across all similarity spaces (phonological and semantic). The degree of cross-correlation between any two similarity spaces can then be computed as the correlation between all matching similarity pairs across the two similarity spaces. Thus, these analyses determine whether the phonological similarity of any two words *Phon*(x,y) correlate with the semantic similarity of those words *Sem*(x,y) (e.g., comparing *Phon*(*dog*,*cat*) with *Sem*(*dog*,*cat*), *Phon*(*dog*,*cog*) with *Sem*(*dog*,*cog*), and so on).

When computing such cross-correlations between the phonological and semantic similarity spaces, Monaghan et al. (in press) found that there was a small positive correlation ($r^2 \approx .002$) indicating that there is systematicity in English sound-meaning correspondences. To make sure that the positive correlation was not a trivial property of the high dimensionality of the similarity spaces, they conducted a set of Monte Carlo analyses. These involved the Mantel (1967) test in which every word's meaning was randomly reassigned (e.g., the meaning for dog might be that of *cat*) and the sound-meaning cross-correlation was then recomputed. This process was repeated 10,000 times, revealing that English words contain more sound-meaning systematicity than would be expected by chance ($p < .0001$). This result was robust across the different phonological and semantic representations. Moreover, to control for possible effects of both inflectional and derivational morphology on form-meaning systematicity, Monaghan et al. further redid their analyses with only monomorphemic versions of the words (*dog* but not *dogs*), and again obtained significant positive correlations. A final set of analyses was conducted to control for potential phonological and/or semantic relatedness due to shared historical origin. For example, words related to the phonaestheme *gl-*, such as *gleam, glitter, glow,* and *glisten*, are proposed to either derive from the Proto-Indo-European root **ghel-*, meaning "to shine, glitter, glow, be warm" (Klein, 1966) or the Old English root **glim-*, meaning "to glow, shine" (OED Online, 2013). To control for relatedness due to common etymology, this set of analyses therefore omitted words with proposed common roots in Old English, Old French, Old Norse, Greek, Latin, Proto-Germanic, or Proto-Indo-European, once more revealing that English incorporates a small but highly significant degree of form-meaning systematicity.

Additional analyses of the contribution of an individual word's form-meaning mapping to the overall systematicity of the vocabulary suggested that the systematicity of English is a property of the language as a whole, and not due to small isolated pockets of words with highly systematic form-meaning mappings. That is, the results cannot be explained by the presence of, in Pinker's terms the "asterisks" of sound symbolism to the general arbitrariness of the sign. Thus, although the very small amount of variance accounted for in the correlation between form and meaning indicate that the mappings between them are largely arbitrary, language nonetheless incorporates a robust amount of systematicity between the sounds of words and their meaning (at least as exemplified by English).

3. The sound of syntax

The results of Monaghan et al. (in press) indicated that the English vocabulary contains a modest but significant amount of systematicity in the mapping between the sound of an individual word and its meaning. Perhaps it is possible to find stronger systematicity within categories of words? Recent corpus analyses provide some initial support for this hypothesis, suggesting that nouns that differ in abstractness (measured in terms of imageability) also tend to differ along several phonological measures, including prosody, phonological neighborhood density, and rates of consonant clustering (Reilly & Kean, 2007). Subsequent psycholinguistic experiments have confirmed that adults are indeed sensitive to such phonological information when making semantic judgments about novel words or reading known words aloud (Reilly, Westbury, Kean & Peelle, 2012). Thus, if abstract and concrete nouns may sound somewhat different from one another, then perhaps there might also be phonological differences between lexical categories of words, such as nouns and verbs?

Phonological cues to how words might be used in a sentence context would likely facilitate the acquisition of syntax. To investigate whether words in this way have the 'sound of syntax' within them, Monaghan, Chater & Christiansen (2005) therefore conducted a series of corpus analyses of child-directed speech to quantify the potential usefulness of phonological cues to lexical categories. More than five million words were extracted from the CHILDES database (MacWhinney, 2000), comprising over one million utterances spoken in the presence of children. Phonological forms and lexical categories for each word were derived from the CELEX database (Baayen et al., 1995) and results reported for the 5,000 most frequent words. As potential cues to lexical categories, Monaghan et al. used 16 different phonological properties (listed in Table 1) that have been proposed to be useful for separating nouns from verbs (and function words from content words). Because each cue is probabilistic, and therefore unreliable when treated in isolation, the 16 cues were combined into a unified phonological representation for each word. Discriminant analyses were then conducted using these representations, resulting in classifications that were significantly better than chance for both nouns (58.5%) and verbs (68.3%)—with an indication that phonological cues may be more useful for discovering verbs than nouns. The advantage of phonological cues for verbs was subsequently confirmed by further analyses in Christiansen and Monaghan (2006).

INSERT TABLE 1 ABOUT HERE

To determine the cross-linguistic validity of these results, Monaghan, Christiansen and Chater (2007) conducted similar analyses for Dutch, French, and Japanese. However, many of the phonological cues used by Monaghan et al. (2005) were specific to English and thus may not work for other languages. Monaghan et al. (2007) therefore generated a set of 53 cross-linguistic phonological cues, including gross-level word cues such as length in phonemes or syllables, consonant cues relating to manner and place of articulation of phonemes in different parts of the word, and vowel cues relating to tongue height and position as well as whether the vowel was reduced. They then conducted analyses of child-directed speech in English, Dutch, French, and Japanese. Using the new cues, they replicated the results of the previous study in terms of correct noun/verb classification in English (16 cues: 63.4% vs. 53 cues: 66.5%). Noun/verb classification using phonological cues was also very good for Dutch (79.6%), French (81.4%) and Japanese (82.2%). The presence of morphology contributed to the classification accuracy for the phonological cues, but was not driving the effects substantially. For instance, analyzing only the monomorphemic words in English resulted in correct classification of 68.0% of nouns and verbs, thus resulting in similar levels of accuracy and statistical significance.

Together, the results of the corpus analyses show that across representatives of three different language genera—Germanic (English, Dutch), Romance (French), and Japanese—child-directed speech contains useful phonological information for distinguishing between nouns and verbs (see also, Kelly, 1992). Crucially, this outcome is not dependent on the specific phonological representations used by Monaghan et al. (2007), as the cross-linguistic results have been replicated using just the initial and final phoneme/mora of a word (Onnis & Christiansen, 2008). More generally the results are consistent with the hypothesis that, as a result of the cultural evolution of language, words contain within them the sound of syntax (Christiansen & Dale, 2004; Christiansen, 2013): nouns and verbs differ in terms of their phonology[2]. Importantly, the specific cues differed considerably across languages, suggesting that each language has recruited its own unique constellation of cues to facilitate acquisition and use.

4. Sound symbolism in language acquisition and use

The corpus analyses surveyed so far suggest that there are systematic correspondences between the sound of a word and its meaning—both for individual words and at the level of lexical categories. But is such sound-meaning systematicity useful for language acquisition and processing? Further corpus analyses by Monaghan et al. (in press) provide support for the idea that systematicity might help children get a foothold in language. Specifically, by exploring the

[2] That the phonological forms of words carry information about their syntactic use as nouns or verbs does not necessarily require the postulation of universal lexical categories. Instead, phonological and distributional cues provide probabilistic information about how words can be used in sentential contexts and this is what is assessed by the corpus analyses reported here.

relationship between a database of age of acquisition norms (Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012) and the sound-meaning systematicity of individual words, they found that early-acquired words tended to be more systematic in their sound-meaning correspondences. Given experimental data showing that 3- to 4-month old infants are able to form cross-modal correspondences between sounds and visual properties of objects—such as between spatial height and angularity with auditory pitch (Walker, Bremner, Mason, Spring, Mattock, Slater, & Johnson, 2010)—the sound-meaning systematicity of early words may thus provide a scaffolding for word learning (see also Miyazaki et al., 2013). Indeed, the cross-modal correspondences of early words may even help young infants learn in the first place that there *are* mappings between sound and meaning (Spector & Maurer, 2009).

As the vocabulary grows across development, Monaghan et al. (in press) found that the form-meaning systematicity of later-acquired words decreases. Although this may seem puzzling, it might reflect computational efficiency considerations resulting from the need to represent a large adult vocabulary. Computational simulations by Gasser (2004) showed that systematic form-meaning mappings were helpful for learning small vocabularies where perceptual representations of words can be kept sufficiently distinct from one another while still allowing for sound symbolic systematicity to exist. As the vocabulary grows, however, it becomes increasingly hard for words with similar meanings to have sufficiently different word forms to avoid confusion because parts of the representational space become saturated with word forms. Eventually, with large vocabularies, arbitrary mappings rather than systematic ones end up resulting in better learning overall as word forms can be more evenly distributed across representational space.

Nonetheless, as the systematicity of individual words declines with increased vocabulary size, group-based systematicity may increase in its potential importance by providing cues to a word's abstract syntactic role in a sentence. Preliminary support for the usefulness of systematic phonological cues to facilitate learning about nouns and verbs comes from connectionist simulations by Reali, Christiansen and Monaghan (2003). Using the 16 phonological cues from Monaghan et al.'s (2005) analyses of English as input representations, they trained Simple Recurrent Networks (Elman, 1990) on a corpus of child-directed speech (Bernstein-Ratner, 1984). These networks thus were provided with both phonological cues as well as distributional information that could be learned from the co-occurrence of words in the input. A second set of networks was provided with distributional information only. This was accomplished by randomizing the phonological cues for each word (e.g., all instances of the word *dog* might be assigned the phonological cues for *walk*), breaking any systematic relationship between phonological cues and lexical categories while maintaining the same input representations.

The simulation results revealed that the networks provided with systematic phonological cues as input were significantly better at learning to predict the next lexical category in a sentence, compared to the networks that learned from distributional information alone. Analyses of the networks' hidden unit activations—essentially their internal state at a particular point in a sentence given previous input—revealed that the networks used the phonological cues to place themselves in 'noun state' when processing nouns and in a separate 'verb state' when processing verbs. Thus, the simulations demonstrated the advantage of systematic sound-meaning mappings for acquisition, especially when it comes to processing novel nouns and verbs. Further corroboration comes from Storkel (2001, 2003) who has shown that preschoolers find it easier to learn novel words when these consist of phonotactically common sound sequences.

To more clearly establish whether children may exploit the systematic relationship between sound and lexical categories as revealed by the corpus analyses, Fitneva, Christiansen and Monaghan (2009) conducted a word learning study to investigate whether children implicitly use phonological information when guessing about the referents of novel words. To create novel words with phonological cues to their use as nouns or verbs, Fitneva et al. used a measure of phonological typicality, originally proposed by Monaghan, Chater & Christiansen (2003). Phonological typicality measures how typical a word's phonology is relative to other words in its lexical category, and reliably reflects the phonological systematicity of nouns and verbs (Monaghan, Christiansen, Farmer & Fitneva, 2010). Thus, what we refer to as 'noun-like' nouns are typical in terms of their phonology of the category of nouns, and likewise 'verb-like' verbs are phonologically typical of other verbs. The distinction between noun-like and verb-like is quite subtle, and not easy to discern. For example, *fact* is a noun-like noun whereas *myth* is a verb-like noun; similarly, *learn* is a verb-like verb, whereas *thrive* is a noun-like verb.

Fitneva et al. (2009) created a set of novel words that were either noun-like or verb-like in their phonology and asked English monolingual second-graders to guess whether these words referred to a picture of an object or a picture of an action. The results indicated that the children were using the phonological typicality of the novel word when making their choices. Interestingly, as predicted by the corpus analyses (Christiansen & Monaghan, 2006), verbs benefitted more from phonological cues than nouns. In a further experiment with second-graders taught in a French immersion program, Fitneva et al. demonstrated that relatively little exposure to language is needed in order for children to use phonological typicality to make guesses about novel words (indeed, just two years of exposure to French in a formal education setting was sufficient).

The results of the word learning study suggest that phonological cues may come into play early in syntax acquisition. Farmer, Christiansen and Monaghan (2006) explored whether the use of such systematic sound-meaning correspondences extends into adulthood. Analyzing an existing database of word naming latencies (Spieler & Balota, 1997), they found that the processing of words presented in isolation is affected by how typical their phonology is relative to their lexical category: noun-like nouns are read aloud faster, as are verb-like verbs. Similarly, Monaghan et al. (2010) analyzed a lexical decision database (Balota, Cortese, Sergent-Marshall, Spieler & Yapp, 2004), revealing that people produce faster responses for words that are phonologically typical of their lexical category. Farmer et al. further showed that the phonological typicality of a word could even affect how easy it is to process in a sentence context. Indeed, for noun/verb homonyms (e.g., *hunts* as in *the bear hunts were terrible...* versus *the bear hunts for food...*), if the continuation of the sentence is incongruent with the phonological typicality of the homonym, then people both experience on-line processing difficulties and have problems understanding the meaning of the sentence.

Together, the results of the human experimental studies indicate that the use of sound-meaning systematicity during acquisition is so important that it becomes a crucial part of the developing language processing system. The systematic phonological properties of words—despite their subtlety—facilitate lexical acquisition and become an intricate part of lexical representations. As consequence, adult language users cannot help but pay attention to phonological cues to syntactic structure when processing language.

5. A division of labor between systematicity and arbitrariness

The results of the corpus analyses and human experiments reviewed in this paper demonstrate that language strikes a delicate balance between arbitrariness and systematicity in form-meaning mappings. The acquisition of the initial vocabulary is facilitated by the systematic relationship between sound and meaning in early-acquired words. At this stage, systematicity enables knowledge about known words to be extrapolated to constrain the meaning individuation of new words. However, as the vocabulary grows, such generalizations become less informative at the individual word level, as arbitrariness increases in later-acquired words. We hypothesize that arbitrariness then comes to facilitate learning the meaning of individual words while obscuring potential similarities between individual word meanings. Instead, systematicity at the level of groups of words comes into play, allowing learners to exploit systematic correspondences between phonological forms and lexical categories when acquiring new words.

The change in the role of phonological cues—from meaning individuation to lexical categories—may also signal a change in the relative usefulness of those cues for learning about nouns and verbs. Thus, whereas Monaghan et al. (in press) found no differences between nouns and verbs in terms of the impact of form-meaning systematicity on age of acquisition, phonological cues to lexical categories appear to work better for verbs than for nouns, as evidenced by both corpus analyses (Christiansen & Monaghan, 2006) and developmental experimentation (Fitneva et al., 2009). Because verbs, in comparison to nouns, appear both to be conceptually harder to learn (e.g., Childers & Tomasello, 2006) and occur in less reliable distributional contexts (Monaghan et al., 2007), phonological cues may be particularly important for the acquisition of verbs (Christiansen & Monaghan, 2006). The structure of the vocabulary may in this way reflect a functional pressure in form-meaning mappings toward facilitating verb learning through phonological cues.

More generally, there appears to be a division of labor between arbitrariness and systematicity in word form-meaning mappings, deriving from opposing pressures from the task of learning the meanings of individual words, on the one hand, and the process of discovering how to use these words syntactically, on the other. Monaghan, Christiansen and Fitneva (2011) present results from computational simulations, human experiments, and corpus analyses indicating that whereas one part of a word's phonological form may have a primarily arbitrary form-meaning mapping to facilitate meaning individuation, another part of the same word tends to incorporate systematicity to assist in the acquisition of lexical category information. In corpus analyses of English and French, for instance, word beginnings were found to carry more information for individuation, whereas word endings supported grammatical-category level discovery (see also Hawkins & Gilligan, 1988; St. Clair, Monaghan, & Ramscar, 2009).

Importantly, we do not see the impact of phonology on the vocabulary as mere vestiges, left over from long bygone iconic resemblances between words and their references. Rather, we construe phonological cues as a very active component of how we acquire new words and use the ones we already know, as exemplified by a steadily increasing number of psycholinguistic studies highlighting the role of sound-category correspondences in language processing (e.g., Farmer et al., 2006; Fitneva et al., 2009; Monaghan et al., 2011; Reilly et al., 2012). These phonological cues also appear to provide constraints on how language changes across time. For example, Kelly (1988) showed that lexical stress differences between nouns and verbs—in English nouns tend to have first-syllable stress, whereas verbs generally have second-syllable stress—affect new derivational uses of such words. Specifically, he found that over the history of English, nouns with second-syllable stress have been more likely than nouns with first-syllable stress to develop verb use, and vice versa for verbs with first-syllable stress. Thus, phonological cues not only influences how we acquire and use our current vocabulary but also appear to shape the vocabularies of future language users.

The ‘big data’ analyses we have discussed here indicate that the structure of the vocabulary reflects multiple competing pressures on language acquisition and use. This division of labor is expressed within the vocabulary both within the structure of single words, but also longitudinally, as the requirements on language processing change with the developing vocabulary. The substantial degree of systematicity evidenced for both individual words and at the level of lexical categories forces us to reappraise the century old assumption about the arbitrariness of the sign. Crucially, for the purpose of this special issue, this kind of insight would not have been possible without ‘big data’ analyses of many kinds of language databases, from etymological dictionaries and age of acquisition norms to child-directed speech corpora and collections of lexical processing latencies. Of course, we are not advocating that corpus analyses should replace standard psycholinguistic experimentation but, rather, that ‘big data’ should be welcomed as an important addition to the methodological toolbox of language scientists in their search for further insight into the nature of language.

Acknowledgments

We are grateful for comments by Stevan Harnard and two anonymous reviewers on a previous version of this paper.

References

Baayen, R. H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H. & Yapp, M.J. (2004). Visual word recognition for single syllable words. *Journal of Experimental Psychology: General, 133*, 283-316.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language, 80,* 290-311.

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language, 11*, 557-578.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: MIT Press.

Boas, F. (1940). *Race, language and culture*. Chicago: University of Chicago Press.

Boring, E.G, (1960). *A history of experimental psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Burnard, L. & Aston, G. (1998). *The BNC handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Bowerman, M. (1974). Learning the structure of causative verbs: A study in the relationship of cognitive, semantic, and syntactic development. *Papers and Reports on Child Language Development, 8,* 142-178.

Braine, M.D.S. (1963). The ontogeny of English phrase structure: The first phase. *Language, 39,* 1-13.

Brants, T. & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia, PA: Linguistic Data Consortium.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Childers, J.B. & Tomasello, M. (2006). *Are nouns easier to learn than verbs? Three experimental studies.* In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 311-335). New York: Oxford University Press.

Christiansen, M.H. (2013). Language has evolved to depend on multiple-cue integration. In R. Botha & M. Everaert (Eds.), *The evolutionary emergence of language: Evidence and Inference* (pp. 42-61). Oxford: Oxford University Press.

Christiansen, M.H. & Dale, R. (2004). The role of learning and development in the evolution of language. A connectionist perspective. In D. Kimbrough Oller & U. Griebel (Eds.), *Evolution of communication systems: A comparative approach. The Vienna Series in Theoretical Biology* (pp. 90-109). Cambridge, MA: MIT Press.

Christiansen, M.H. & Monaghan, P. (2006). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R.M. Golinkoff (Eds.), *Action meets words: How children learn verbs* (pp. 88-107). New York: Oxford University Press.

Darwin, C. (1877). A biographical sketch of an infant. *Mind, 2*, 285-294.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*, 447-464.

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

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61-74.

Eco, U. (1995). *The search for the perfect language*. London: Blackwell.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Farmer, T.A., Christiansen, M.H. & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences, 103*, 12203-12208.

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K. (Ed.), *Encyclopedia of language and linguistics* (2nd Ed., pp. 665-670). Oxford: Elsevier.

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

Gasser, M. (2004). The origins of arbitrariness of language. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 434-439). Mahwah, NJ: Erlbaum.

Gasser, M., Sethuraman, N. & Hockema, S. (2011). Iconicity in expressives: An empirical investigation. In J. Newman & S. Rice (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 163–180). Stanford, CA: Center for the Study of Language and Information.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Hamilton, E. & Cairns, H. (Eds.) (1961). *Plato: The collected dialogues*. Princeton, NJ: Princeton University Press.

Harris, Z.S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hawkins, J. A., & Gilligan, G. (1988). Prefixing and suffixing universals in relation to basic word order. *Lingua, 74*, 219–259.

Hockett, C. F. (1954). Two models of grammatical description. *Word, 10*, 210-234.

Hockett, C. F. (1960). The origin of speech. *Scientific American, 203*, 89-96.

Huettig, F., Quinlan, P., McDonald, S. & Altmann, G.T.M. (2006). Word co-occurrence statistics predict language-mediated eye movements in the visual world. *Acta Psychologica, 121*, 65-80.

Ingram, D. (1989). *First language acquisition: Method, description, and explanation*. Cambridge, NY: Cambridge University Press.

Kading, J. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: Privately published.

Kelly, M.H. (1988). Phonological biases in grammatical category shifts. *Journal of Memory and Language*, *27*, 343-358.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*, 349-364.

Klein, E. (1966). *A comprehensive etymological dictionary of the English language: Dealing with the origin of words and their sense development thus illustrating the history of civilization and culture.* Amsterdam: Elsevier.

Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavioral Research Methods, 44*, 978-990.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (Third Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. & Snow, C. (1985). The child language exchange system. *Journal of Child Language, 12,* 271-296.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209-220.

McEnery, T. & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Miller, G. A. (2013). “WordNet—About us”, WordNet. Retrieved from http://wordnet.princeton.edu.

Miyazaki, M., Hidaka, S., Imai, M., Yeung, H, H., Kantartzis, K., Okada, H. & Kita, S. (2013). The facilitatory role of sound symbolism in infant word learning. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3080-3085). Austin, TX: Cognitive Science Society.

Monaghan, P., Chater, N. & Christiansen, M.H. (2003). Inequality between the classes: Phonological and distributional typicality as predictors of lexical processing. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 810-815). Mahwah, NJ: Lawrence Erlbaum.

Monaghan, P., Chater, N. & Christiansen, M.H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96*, 143-182.

Monaghan, P. & Christiansen, M.H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (TILAR Series) (pp. 139-163). Amsterdam: John Benjamins.

Monaghan, P., Christiansen, M.H. & Chater, N. (2007). The Phonological-Distributional Coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259-305.

Monaghan, P., Christiansen, M.H., Farmer, T.A. & Fitneva, S.A. (2010). Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon, 5*, 281-299.

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

Monaghan, P., Shillcock, R.C., Christiansen, M.H. & Kirby, S. (in press). How arbitrary is language? *Philosophical Transactions of the Royal Society B*.

Morgan, J.L. & Demuth, K. (Eds.). (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum.

OED Online (2013). Oxford: Oxford University Press (accessed July 05, 2013).

Olshausen, B.A. & Field, D.J. (2005). How close are we to understanding V1? *Neural Computation, 17*, 1665-1699.

Onnis, L. & Christiansen, M.H. (2008). Lexical categories at the edge of the word. *Cognitive Science, 32*, 184-221.

Quirk, R. (1960). Towards a description of English Usage. *Transactions of the Philological Society*, *59,* 40-61.

Perniss, P., Thompson, R. L. & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227. doi: 10.3389/fpsyg.2010.00227.

Pinker, S. (1999). *Words and rules*. New York: Basic Books.

Reali, F., Christiansen, M.H. & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration. In

Proceedings of the 25th Annual Conference of the Cognitive Science Society (pp. 970-975). Mahwah, NJ: Lawrence Erlbaum.

Reilly, J. & Kean, J. (2007). Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive Science, 31*, 1-12.

Reilly, J., Westbury, C., Kean, J. & Peelle, J.E. (2012). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PLoS ONE* 7(8): e42286.

Riordan, B. & Jones, M.N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3*, 303-345.

Schmidtke, D.S., Conrad, M. & Jacobs, A.M. (2014). Phonological iconicity. *Frontiers in Psychology,* 5:80. doi: 10.3389/fpsyg.2014.00080.

Spector, F. & Maurer, D. (2009). Synesthesia: A new approach to understanding the development of perception. *Developmental Psychology, 45*, 175-189.

Spieler, D.H. & Balota, D.A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411-416.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*, 1317-1329.

Storkel, H.L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44*, 1321-1337.

Storkel, H.L. (2003). Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research, 46*, 1312-1323.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge: Cambridge University Press.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21,* 21–25.

Table 1: Examples of the 16 Phonological Cues used by Monaghan, Chater & Christiansen (2005)

Phonological Cue	*penguin*	*cat*	*dog*
Word level			
Length in phonemes	7	3	3
Length in syllables	2	1	1
Presence of stress	1	1	1
Syllable position of stress	1	1	1
Syllable level			
Number of consonants in word onset	1	1	1
Proportion of phonemes that are consonants	0.71	0.67	0.67
Proportion of syllables containing reduced vowel	0	0	0
Reduced 1st vowel	0	0	0
-ed inflection	0	0	0
Phoneme level			
Proportion of consonants that are coronal	0.2	0,5	0.5
Initial /ð/	0	0	0
Final voicing	1	0	1
Proportion of consonants that are nasals	0.4	0	0
Position of stressed vowel (see following)	1	1	3
Position of vowels (from 1 = front, to 3 = back)	1.25	1	3
Height of vowels (from 0 = close, to 3 = open)	1.25	2.5	3