Available online at www.sciencedirect.com
ScienceDirect
Cognitive Psychology 55 (2007) 259–305
Cognitive Psychology
www.elsevier.com/locate/cogpsych

The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition ☆

Padraic Monaghan [a,*], Morten H. Christiansen [b], Nick Chater [c]

[a] *Department of Psychology, University of York, York, YO10 5DD, UK*
[b] *Department of Psychology, Cornell University, Ithaca, NY 14853, USA*
[c] *Department of Psychology, University College London, Gower Street, London, WC1E 6BT, UK*

Accepted 19 December 2006
Available online 8 February 2007

Abstract

Several phonological and prosodic properties of words have been shown to relate to differences between grammatical categories. Distributional information about grammatical categories is also a rich source in the child's language environment. In this paper we hypothesise that such cues operate in tandem for developing the child's knowledge about grammatical categories. We term this the Phonological-Distributional Coherence Hypothesis (PDCH). We tested the PDCH by analysing phonological and distributional information in distinguishing open from closed class words and nouns from verbs in four languages: English, Dutch, French, and Japanese. We found an interaction between phonological and distributional cues for all four languages indicating that when distributional cues were less reliable, phonological cues were stronger. This provides converging evidence that language is structured such that language learning benefits from the integration of information about category from contextual and sound-based sources, and that the child's language environment is less impoverished than we might suspect.

☆ This research was supported by Human Frontiers of Science Program Grant RGP0177/2001-B. We are grateful to Marjolein Merkx of the University of Warwick for assistance in preparing the Dutch corpus, Luca Onnis of Cornell University for assistance in preparing the French corpus, and Mikihiro Tanaka of Edinburgh University and Yuki Kamide of Dundee University for assistance with the Japanese corpus and analyses.

* Corresponding author. Fax: +44 1904 433181.
E-mail addresses: P.Monaghan@psych.york.ac.uk, pjm21@york.ac.uk (P. Monaghan).

0010-0285/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2006.12.001

Keywords: Language acquisition; Syntactic bootstrapping; Phonology; Distributional information; Poverty of the stimulus

1. Introduction

Learning grammatical categories is essential in order for the child to develop an understanding of the relationships between sounds in a spoken sentence and objects and actions in the world around them (Gentner, 1982). Knowledge of the patterns determining which words can relate to objects, which to actions, and which modify the relationships between these objects and actions is an imperative in language development (Pinker, 1984).

One view of this acquisition process is that the child has innate constraints that facilitate this development. Some theorists argue that these constraints encode a complete grammar of human natural language, aside from a finite set of parametric variations that define the structural differences between languages (e.g., Baker, 2001; Chomsky, 1965, 1981; Crain & Lillo-Martin, 1999). From this perspective, the entire grammatical machinery of natural language is innate—and hence the set of possible syntactic categories, including nouns, verbs, adjectives, and so on, must similarly be innate. The child's task, under this view, is to learn which words belong to which syntactic categories.

Alternatively, Pinker's (1984) semantic bootstrapping hypothesis predicts rather that certain *semantic* referents are innately specified, and reflected in the surface properties of the language in terms of distributional co-occurrence information. Thus, for the noun/verb distinction, the child has innately specified information in terms of nouns referring to objects, and verbs referring to actions. These semantic referents then constrain the child's search for relevant correlations in the language to which she is exposed, and also, according to Pinker, provide an explanation for why such correlations between surface distributional properties and semantic features occur in natural languages (e.g., that nouns and verbs occur in different distributional contexts). Pinker (1984, p.43) states "it [semantic bootstrapping] claims that children always give priority to distributionally based analyses, and is intended to explain how the child knows which distributional contexts are the relevant ones to examine." Whether the innately specified language structure is syntactic or semantic, the child also faces a further task: learning which grammatical categories are realized in the language, given that not all possible categories occur in all languages (e.g., Croft, 2003; Dixon, 1977). According to some recent, and influential, linguistic analyses, the child's task, under the nativist position, may be more complex than previously assumed due to the extraordinary variety of fine-grained syntactic categories in natural language (e.g., Culicover, 1999).

The view that some knowledge about the language, or grammatical categories of the language, is innately specified is typically based, at least in part, on the assumption that there is insufficient evidence in the child's language environment to enable these properties to be learned from the language itself. That is, nativist viewpoints concerning the origins of syntactic categories typically rely, to some degree, on arguments from the "poverty of the stimulus" (e.g., Chomsky, 1980; though see Pullum & Scholz, 2002). Under the semantic bootstrapping account, for instance, it is claimed that learning the correlations between grammatical categories and distributional information of their usage ought to be impossible as the search for correlations is too unconstrained. Yet, a study by Gerken, Wilson, and Lewis (2005) demonstrated that such learning *is* possible in children younger than two

years of age when no semantic, referential information was available. Their participants learned to distinguish grammatical from ungrammatical gender-marked nouns after brief exposure to examples of the language, but only under conditions where there were two partially overlapping phonological cues to the grammatical distinction.

Such category learning from correlational information alone has only been shown in relatively restricted domains. The search space for correlations in natural languages is vastly greater than in artificial language studies, and so, as Pinker (1984, p.50) notes, "the properties that the child can detect in the input—such as the serial positions and adjacency and co-occurrence relations among words—are in general linguistically irrelevant." If learning from natural languages is unconstrained from a source other than distributional information, then the child may well learn correlations that are inconsistent with the language. Thus, from *John eats meat, John eats slowly*, and *The meat is good*, the child incorrectly infers that *The meat is slowly* is also an acceptable expression, though see Cartwright and Brent (1997) for a distributionally-based solution to this problem. Thus it is possible that participants in artificial language learning studies with no referential information available learn correlations that are consistent but errorful, but without testing sequences that are consistent but illegal in artificial languages the extent to which distributional learning alone is constrained has not been fully established.

There are, however, alternative sources of constraints on learning the correlations from distributional information, due to the relationship between prosodic and phonological properties of speech and syntactic structure (Morgan & Newport, 1981). For example, Cooper and Paccia-Cooper (1980) indicated that in natural speech phrase structure was related to prosodic properties, though prosodic cues were not found to distinguish between noun phrases and verb phrases. Additionally, there are correspondences between grammatical categories and phonological properties in English (Kelly, 1992; Monaghan, Chater, & Christiansen, 2005), as well as in gender as noted in Gerken et al. (2005). However, for the phonological and prosodic constraints to qualify potentially as an essential constraint on learning grammatical categories, these cross-modal correlations have to be observed across all languages. In this paper, we argue that the child's language environment is not as impoverished as has been assumed, if one considers a variety of sources of information in the speech signal other than only information about word identity and word order. We make the case that multiple cues that are available to the child in language learning can contribute to the development of accurate and useful grammatical categories, and that general learning mechanisms based on these multiple sources may well be adequate for beginning the process of category development in the child. Our argument will be built around the Phonological-Distributional Coherence Hypothesis—that phonological and distributional properties of words interact in a way that provide useful, and perhaps ultimately sufficient, constraints for developing grammatical categories in language acquisition.

What sources of information, then, may the child utilize in order to construct this sense of grammatical categories, and membership of particular words within those categories? Studies of the properties of the English language have indicated the importance of multiple cues that signify the grammatical category of the word (Durieux & Gillis, 2001; Fernald & McRoberts, 1996; Finch & Chater, 1992; Fisher & Tokura, 1996; Gerken, 2001; Höhle, Weissenborn, Schmitz, & Ischebeck, 2001; Kelly, 1992; Mintz, 2003; Morgan & Demuth, 1996; Onnis & Christiansen, in press; Redington, Chater, & Finch, 1998). Relatedly, artificial language learning experiments have indicated that the conjunction of such multiple

cues is valuable, and at times necessary, for supporting learning of language structure (Braine, 1987; Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Mintz, 2002; Monaghan et al., 2005; Morgan, Meier, & Newport, 1987; Onnis, Monaghan, Richmond, & Chater, 2005).

Studies of cues that are effective in predicting the grammatical category of a word have focused on properties that are either internal or external to the word. External cues are those that determine the word's usage from its context, such as distributional information[1]—the position of the word in relation to other words in the utterance (Bloomfield, 1933; Campbell & Besner, 1981; Cartwright & Brent, 1997; Durieux & Gillis, 2001; Harris, 1954; Maratsos & Chalkley, 1980; Mintz, 2003; Redington et al., 1998)—or deictic, gestural, and semantic information (e.g., Bowerman, 1973; Tomasello, 2003).

In contrast, information can also be found within the word itself, and concerns phonological or prosodic information—the sound of the word and its correspondence to different grammatical categories (Brooks et al., 1993; Cassidy & Kelly, 1991, 2001; Cutler, 1993; Cutler & Carter, 1987). In this paper, we focus on one type of external cue—distributional information—and one type of internal cue, the phonological properties of the word. These types of cue can be quantitatively assessed through the use of child-directed speech corpora. We next review studies of these cue types in language development. Most studies focus on English, but we report the rare cases where studies have taken a cross-linguistic perspective.

2. Phonological cues to syntactic categories

Kelly (1992) reviewed a range of phonological cues that have been proposed as corresponding to particular syntactic categories in English. Several cues were related to distinguishing open from closed class words for example, open class words tend to have longer syllable duration, and are more likely to contain consonant clusters (Morgan, Shi, & Allopenna, 1996). Shi (1995), reported in Shi, Morgan, and Allopenna (1998) analysed English child-directed speech and found that closed class words were more likely to contain centralized vowels, and were less likely to have consonants in the word onset.

Other cues were found to distinguish nouns from verbs. For example, in English, disyllabic nouns are more likely to have first-syllable stress whereas disyllabic verbs are more likely to have second-syllable stress (Kelly, 1992). Swingley (2005) suggested that stress was important for segmenting speech into word forms that could then be clustered according to their distributional characteristics into syntactic categories. Nouns are also longer than verbs in English (Cassidy & Kelly, 1991), and the use of this cue for vocabulary acquisition was explicitly tested in experiments where children were given either one or three syllable nonwords and asked to guess whether the nonword referred to an object or an action (Cassidy & Kelly, 2001). Longer nonwords were more likely to be used as nouns (referring to objects), and shorter nonwords were more likely to be used as verbs (with reference to actions). Additionally, nouns were found to have a greater probability of containing low vowels and nasal consonants than verbs (Kelly, 1992).

[1] We use the term distributional information to refer to information derived from co-occurrence with other words. The phonological information can also be seen as distributional in that the cues are useful because of their different distributions across the word classes. We persevere with the terms phonological and distributional to align them with other studies in the literature on co-occurrence information (Mintz, 2003; Redington et al., 1998).

Durieux and Gillis (2001) tested several cues reported by Kelly (1996) for their effectiveness in classifying 5000 words taken from the CELEX database (Baayen, Pipenbrock, & Gulikers, 1995). The cues assessed were stress position, vowel height for each syllable, presence of nasal consonants, and number of phonemes per syllable. For the noun/verb distinction, an instance based learning model learned to classify almost 68% of words correctly. An encoding of phonemes in onset, nucleus and coda positions for each word was also performed, to see if particular phonemes in certain positions distinguished nouns from verbs. In this case, the analysis resulted in 74% correct classifications, which rose to 78% when stress position was also included. An additional assessment on distinguishing nouns, verbs, adjectives, adverbs, and words that were ambiguous between these four syntactic categories was also performed. In this case, almost 67% of words were correctly classified using the phoneme by position encoding. Combined cues, therefore, contributed significantly towards distinguishing open class categories in English.

Monaghan et al. (2005) analysed the 5000 most frequent words from a child-directed speech corpus, testing a range of 16 phonological cues, involving either properties of the word as a whole, the syllable, or the phoneme. Word-level cues related to inflections on the word, such as the pronunciation of-*ed* at the end of a word (Marchand, 1969), the position of stress within the word (Cutler & Carter, 1987; Morgan & Saffran, 1995), or the number of phonemes or syllables in the word (Kelly, 1988). Syllable-level cues referred to size of consonant clusters in the syllable (Cutler, 1993), or the types of phoneme that occur across the syllable. Phoneme-level cues referred to properties of particular phonemes found within the word, such as the proportion of coronal consonants in the word, or the height or position of vowels within the word (e.g., Sereno & Jongman, 1990), all purportedly important cues for distinguishing different syntactic categories.

Research on phonological properties in languages other than English are scarce, though with a few notable exceptions. Shi et al. (1998) performed a detailed cross-linguistic analysis of the auditory properties of child-directed speech in Mandarin and Turkish. They assessed two mother-child dyads for each language, and analysed these for auditory properties that distinguished open from closed class words. For the Mandarin speakers, 98 words were analysed from the first mother, and 77 words from the second. For the Turkish dyads, 100 open and 100 closed class words were analysed for each speaker. Several cues were found that were indicators of grammatical category across the two languages, and that were also found in previous research by the same group on English (Morgan et al., 1996). Closed class words had fewer syllables, shorter vowel duration, fewer syllable codas, fewer vowel diphthongs (in Mandarin), less vowel harmony (Turkish), and less amplitude change. Though individual cues were found to be unreliable for predicting grammatical category, when combined with information about utterance position and frequency as the input to self-organising neural network models, they were found to predict distinctions in approximately 80% of words of which around 90% were accurately classified in Mandarin, and 80-85% in Turkish. The auditory and phonological analyses of these different languages indicate that phonological cues provide a potentially useful source to aid in determining distinctions between grammatical categories, and the generality of the findings suggest that such information may well contribute towards beginning the process of syntactic bootstrapping in language acquisition.

Durieux and Gillis (2001) discovered that the same cues described by Kelly (1996) found to be computationally useful for categorising in English proved even more effective in categorising Dutch. For the noun/verb distinction, over 75% of Dutch words were correctly

classified. In the phoneme by position encoding described above for their English analysis, classification rose to 82%, and 83% if stress was also included. For distinctions between all open class categories, performance was again slightly higher than for English, with 71% correct classification. Their results indicate that not only are phonological cues important for distinguishing syntactic categories in languages other than English, but that the very same cues may be informative for different languages. However, the position may be different for languages that share less in common than the Germanic languages of English and Dutch. Below, we report studies of languages that come from different families to test the extent to which phonological cues are effective in determining syntactic category. Fisher and Tokura (1996), for instance, found that prosodic cues were effective in both English and Japanese for signalling syntactic category.

The phonological cues that have been discovered for English originate from many different sources. Many of these cues are linguistically informed, for example, the stress distinction in English between the noun 'subject and the verb sub'ject, or the pronunciation of the-ed inflection for verbs compared to adjectives (cf. the monosyllabic verb *learned* and the disyllabic adjective *learn-ed*). However, other cues result from general phonological properties of the language that correspond to particular syntactic categories, for example, the finding that vowel position and vowel height distinguish nouns from verbs (Kelly, 1992), and these cues can be discovered by an empirical search of phonological properties that align with syntactic categories. This is the approach we adopt in this paper for a range of different languages. We describe this approach in more detail below. First, we review studies of distributional cues to syntactic category, and then motivate our hypothesis about the interplay of phonological and distributional cues.

3. Distributional cues to syntactic categories

The context of a word with respect to other words in the same sentence provides strong cues about the category of a word in English. Redington et al. (1998) assessed the extent to which the distributional context of words of the same category was similar. They counted the occurrence of 150 frequent words either preceding or following each target word in a corpus of child-directed speech. The resulting co-occurrence vectors for words were compared in terms of their similarity and subjected to a cluster analysis. Syntactic categories of words were taken from the CELEX corpus (Baayen et al., 1995), according to the most frequent usage for each word, and the clusters of words were assessed in terms of whether they contained words of the same category. They found that words of the same syntactic category tended to cluster together, indicating that distributional information was sufficient to produce groupings of words that conformed to labeled syntactic categories (see also Cartwright & Brent, 1997; Harris, 1954; Mintz, Newport, & Bever, 2002; Wolff, 1988). Such methods were found to be generalisable to other languages, such as Chinese (Redington et al., 1995).

The analyses of Redington et al. (1998) provided evidence to show that distributional information was potentially of great value in learning syntactic categories. Yet, the precise form of distributional information that is useful and usable by the child has not yet been determined (Monaghan & Christiansen, 2004). Fries (1952) noted that words only from one category can be used in certain contexts. For example, any word that can be used in the gap "you—to" is a verb. Mintz (2002) showed in an artificial language learning experiment that nonwords occurring in such context "frames", where the preceding and succeeding

word were fixed, could be grouped together. In a study of corpora of child-directed speech, Mintz (2003) found that high-frequency frames in the corpus could predict the category of the intervening word with high accuracy.

Monaghan and Christiansen (2004) found that the preceding word predicted the category of the next word with good accuracy, and also categorized more than four times as many words as taking the preceding and succeeding word frames. Valian and Coulson (1988) produced categorization in an artificial language learning task based just on high-frequency preceding words, and in a large-scale analysis of child-directed speech, Monaghan et al. (2005) assessed the 20 most frequent words from a large corpus of child-directed speech and measured the association between each of these words and the following word. If a word often occurred after one of the target words then the association was high, if the word seldom occurred in this local context then the association was low. Using these very local cues, discriminant analysis resulted in accurate classification of nouns and verbs, and open and closed class words. We use this approach in the cross-linguistic analyses presented in this paper as these local bigram cues provide a good indication of potentially valuable distributional information to syntactic category in English, and they can also be generalized across languages. We describe the generation of these distributional cues in more detail in the first experiment. Categorization based on these cues is likely to underestimate the potential information available in distributional information. For example, simple recurrent networks using word order information perform better in syntactic categorization of words than the discriminant analyses based on co-occurrence bigrams (Reali, Christiansen, & Monaghan, 2003). Yet such information is likely to be within the realm of infant learning. We now provide some justification for our claims for the serendipitous arrangement of phonological and distributional cues related to syntactic category.

4. The Phonological-Distributional Coherence Hypothesis

Phonological cues may be particularly important for learning the category of words when there is little other information available about the word. Noun gender, for instance, has been proposed as a distinction for which phonological cues are crucial for learning (Braine, 1987) as there is an absence of semantic or contextual cues for this category. In artificial language learning studies, such additional phonological cues appear to be necessary for category learning to proceed (Braine et al., 1990; Brooks et al., 1993; Frigo & McDonald, 1998). In French, nouns with phonological cues typical of the gender were identified more quickly (Desrochers, Paivio, & Desrochers, 1989), and correspondence between phonology and gender has also been found in other languages (e.g., German: Mills, 1986; Italian: Bates, Devescovi, Pizzamiglio, D'Amico, & Hernandez, 1995; and Hausa: Corbett, 1991). Even in English, female names are distinct from male names in terms of phonological form (Cassidy, Kelly, & Sharoni, 1999).

Monaghan et al. (2005) discovered an interaction between the usefulness of phonological cues and distributional cues for words of different frequencies. For the highest frequency words, distributional information was especially abundant, but phonological cues did not match categories so closely. However, for lower-frequency words distributional information is less reliable and for these words the phonological cues were most effective. Our approach develops the suggestion of Braine (1987) that, in order for words to be effectively categorised, there must be some phonological coherence to the word set. Additionally, we propose that when distributional information is present the phonological cues are

less crucial and some shifting of these cues related to category can occur. However, when distributional information is weaker then the coherence of phonological cues within the category becomes more important.

A similar interaction between the value of phonological and distributional cues has been found within English for nouns and verbs. Christiansen and Monaghan (2006) compared the potential of each cue type for accurately classifying nouns or verbs. In this study, we found that for nouns distributional and phonological cues were equally useful, whereas for verbs, phonological cues were more reliable than distributional cues. Further, the phonological cues contributed to greater accuracy of classifying verbs than nouns, whereas the opposite effect emerged for the distributional cues. When combined, the cues resulted in similar accuracy of classification for both nouns and verbs. Verbs, with more variation in the contexts in which they can occur, require greater consistency in the phonological cues that relate to the word's category.

Artificial language learning studies have shown that, not only is phonological information useful for learning of grammatical categories and the structure of artificial grammars (Newport & Aslin, 2004; Onnis et al., 2005; Perruchet, Tyler, Galland, & Peereman, 2004), but indeed may be essential in order for category learning to take place effectively, particularly when the structure of the language is complex, as in natural languages. If the co-occurrence of phonological and distributional cues to mark grammatical category, as found in English, is an adaptive property of the language in order to make it more easily learnable, then such a pattern ought to be observed in other languages. We term this the Phonological-Distributional Coherence Hypothesis (PDCH), which predicts that there will be correspondence between phonological properties of words and their grammatical category. Notice, of course, that such an adaptive account does not require a "designer" of the lexicon. Rather, lexical forms that are more easily learnable will be more readily acquired by the next generation of language users; those which are difficult to learn will rapidly be extinguished. This type of adaptive explanation is widespread in the study of language change (e.g., Briscoe, 2002; Christiansen & Ellefson, 2002; Hopper & Traugott, 1993).

We report below analyses of phonological and distributional cues in a language similar to English—Dutch—to see whether the properties of English generalize to another Germanic language. Local co-occurrence information may vary between Dutch and English as typically verbs occur initially in verb phrases, whereas in Dutch, verbs occur immediately after the initial phrase in main clauses and clause-finally in subordinate clauses. We also report analyses of languages distinct from English in other ways: French, which has different prosodic properties to English (Cutler, Mehler, Norris, & Segui, 1992), and Japanese. Japanese has flexible word order, where verbs are clause-final but its arguments can occur in varied order, which means that the distributional information about the word may be less reliable. Japanese also has very different phonological structure, based around the mora instead of the syllable, with morae composed of at most one consonant and one vowel. Thus, the possibility of complex phonological properties such as consonant clusters to signify grammatical categories is reduced in this language. We take Japanese to be a strong test of the PDCH due to its word-order and phonological differences to English.

If any one of these languages demonstrates no close correspondence between grammatical category and phonological properties of words then that indicates that the learnability of the language is not dependent upon such multiple-cue integration, and the PDCH, at least in its present form, is disconfirmed. If all four languages demonstrate a similar correspondence as found in English, then that provides converging evidence that such

phonological coherence within grammatical categories may be a widespread and perhaps universal property of languages, and that it may be an important, even essential, feature of the language to facilitate acquisition.

The PDCH is founded upon the principle that language is more easily learnable when coherence between information sources and category is present. A corollary of this principle is that when one source of information is weaker at determining the word's category, then other cues will be more emphatic. This pattern is found in English (Christiansen & Monaghan, 2006; Monaghan et al., 2005), but does it apply in the other three languages? We predict that words that are ineffectively classified into their grammatical category using distributional cues will be effectively classified using the phonological cues. We therefore test, across the four different languages whether the overall value of distributional information is balanced by the phonological cues.[2]

The experiments we now present report data from analyses of potential phonological and distributional cues for learning syntactic categories. Experiment 1 assessed whether phonological cues, at the word-, syllable-, and phoneme-level related to grammatical category in English, Dutch, French, and Japanese. We also include analyses of English here to ensure that this cue-search approach results in similar performance to the use of linguistically-informed phonological cues. Experiment 2 assessed whether high-frequency words immediately preceding or succeeding the target word were good reflections of word category across the different languages. Experiment 3 tested the relative contribution of distributional and phonological cues for determining the syntactic category of words in these different languages. Experiment 4 repeated the combined analyses on a part-of-speech tagged corpus, investigating the effect of grammatical category ambiguity on the results. Finally, Experiment 5 tested the extent to which the PDCH was maintained when inflectional and derivational morphology was removed from the language.

5. Experiment 1: cross-linguistic analyses of phonological cues

5.1. Method

5.1.1. Corpus preparation

For the English corpus, we selected all the adult speech spoken in the presence of the child—so this incorporated all adult-to-adult and adult-to-child speech from the CHILDES corpus (MacWhinney, 2000). It was not possible to distinguish adult-adult and adult-child speech from the corpora, and so we included all adult speech spoken in the presence of children. We assumed that all these utterances provide potentially useful information to the child in learning their first language, though the differences between adult-adult and adult-child speech are well-attested (Bernstein Ratner & Rooney, 2001). Pauses and turn-taking were marked as utterance boundaries resulting in 5,436,855 words in 1,369,574 separate utterances. The phonology and most common syntactic category for each word were taken from the CELEX database (Baayen et al., 1995; Roach & Hartman, 1997). Words with alternative pronunciations or category were assigned their most common usage. We counted the frequency of words in the corpus, and all words in the most frequent 1000 words that did not occur in CELEX were hand-coded for pronunciation and category by a native speaker of English.

[2] We were not able not pursue prosodic cues across the languages, as our corpora did not incorporate this information for every language.

The Dutch corpus was comprised of the 915,302 words of adult-to-adult and adult-to-child speech from the CHILDES Dutch corpus. Utterance boundaries were marked in the same way as for English, resulting in 177,510 separate utterances. The most frequent grammatical category and pronunciation was taken from CELEX, and words that did not occur in the CELEX corpus were hand-coded by a native Dutch speaker.

The French corpus was generated in a similar manner. All child-directed and between-adult speech from the CHILDES French corpus was taken, which resulted in 379,402 words in 79,012 utterances. Pronunciation and class was taken from the LEXIQUE database (New, Pallier, Ferrand, & Matos, 2001), with words from the most frequent 1000 words that did not occur in LEXIQUE hand-coded by a native speaker of French.

The Japanese corpus was formed from all child-directed and between-adult speech in the portion of the Japanese CHILDES database that was transcribed into romaji, with utterance boundaries marked in the same way as for English, Dutch, and French. There were 358,401 words in 138,171 utterances. Phonological form and grammatical category was taken from the Japanese CALLHOME corpus (Canavan & Zipperlen, 1996), with the most frequent 1000 words that did not occur in the CALLHOME corpus coded for most frequent grammatical category by a native Japanese speaker consulting examples from the corpus. For all words not in the CALLHOME corpus, phonology was generated by applying orthography-to-phonology pronunciation rules (McCawley, 1968; Vance, 1987).

5.1.2. Cue generation

For each word, we computed a set of cues based on the phonology of the word in order to assess whether these cues were differently distributed across the grammatical categories. At the word level, we measured the number of syllables (or morae in Japanese), the number of phonemes in each word, and the proportion of phonemes in the word that were consonants (syllabic complexity) and the proportion that were vowels (vowel density)—these two latter measures are related. At the phoneme level, we measured the proportion of consonants with particular manner and place features, and the average height and position of vowels. These measures were made across the whole word, just the first syllable, or just the first phoneme. This was because the beginnings of words have been suggested to be more important in reflecting grammatical category than medial or final phonemes (Durieux & Gillis, 2001; Kelly, 1992). In addition, we determined the proportion of vowels that were reduced (occurring as/ə/) across the word, a distinction reflecting the open/closed class distinction in English (Morgan et al., 1996). In all, there were 53 phonological cues measured. Table 1 shows the entire list of cues. For certain languages, certain cues were not relevant. For instance, English has dental consonants (/θ/,/ð/) whereas Dutch, French, and Japanese do not. French has nasal vowels, but English, Dutch, and Japanese do not. Japanese has flap consonants, but English, Dutch, and French do not.

In order to assess the validity of phonological cues in distinguishing syntactic categories, we compared the means for open and closed class words in the 1000 most frequent words from each language corpus. Open class words were nouns, adjectives, verbs, and adverbs. Articles, pronouns, conjunctions, and prepositions constituted the closed class words we examined. We omitted words classified as proper nouns, numerals, interjections, and contractions (e.g., *I'd*, *would've*). We were also interested in finer discriminations within the open class words, and so we compared means for nouns and verbs. The type and token frequency for open and closed class words and nouns and verbs are shown in Table 2 for each

Table 1
Phonological cues tested

Cue place	Cue
Whole word	Length in syllables/morae
	Length in phonemes
	Syllabic complexity
	Coronals in word
	Voiced consonants in word
	Plosives in word
	Nasals in word
	Trills in word
	Fricatives in word
	Approximants in word
	Flaps in word
	Bilabials in word
	Velars in word
	Alveolars in word
	Palatals in word
	Labials in word
	Uvulars in word
	Glottals in word
	Dentals in word
Onset	Plosives in onset
	Nasals in onset
	Trills in onset
	Fricatives in onset
	Approximants in onset
	Flaps in onset
	Bilabials in onset
	Labials in onset
	Alveolars in onset
	Velars in onset
	Uvulars in onset
	Glottals in onset
	Dentals in onset
	Voiced consonants in onset
First consonant	Plosives in first consonant
	Nasals in first
	Trills in first
	Fricatives in first
	Approximants in first
	Flaps in first
	Bilabials in first
	Labials in first
	Alveolars in first
	Velars in first
	Uvulars in first
	Glottals in first
	Dentals in first
	Voiced consonant first
Vowels	Vowel density
	Reduced vowels
	Vowel position

(*continued on next page*)

Table 1 (*continued*)

Cue place	Cue
	Vowel height
	Rounded vowels
	Nasal vowels

Table 2
Types and tokens of open and closed class words, nouns and verbs for the four languages

Language	Open		Closed		Noun		Verb	
	Type	Token (%)	Type	Token (%)	Type	Token (%)	Type	Token (%)
English	757	43.0	97	33.3	380	11.6	218	18.8
Dutch	761	48.2	94	37.2	306	7.1	238	19.8
French	853	43.0	107	50.5	382	9.0	328	21.9
Japanese	790	37.9	87	40.9	415	17.0	276	14.7

language. Type frequency indicates the number of each category from the 1000 words, and token frequency indicates the proportion of the whole corpus represented by these words.

5.2. Results

5.2.1. Open and closed class words

Tables 3–6 show the means for open and closed class words for significant cues in English, Dutch, French, and Japanese, respectively. *T*-tests were Bonferroni-corrected for

Table 3
Phonological cues significant for open/closed class words in English

Cue	English		
	Open	Closed	*t*
Length in phonemes	3.95	3.41	3.74**
Syllabic complexity	0.61	0.57	3.01[a]
Voiced consonants in word	0.60	0.74	−3.75**
Plosives in word	0.42	0.18	6.56***
Fricatives in word	0.24	0.38	−3.44*
Velars in word	0.15	0.03	8.53***
Dentals in word	0.01	0.12	−3.97**
Plosives in onset	0.25	0.10	4.99***
Fricatives in onset	0.13	0.29	−3.93**
Velars in onset	0.08	0.01	9.77***
Dentals in onset	0.01	0.09	−3.20*
Voiced consonants in onset	0.32	0.51	−4.91***
Plosives in first consonant	0.40	0.08	9.98***
Bilabials in first	0.22	0.11	3.14*
Alveolars in first	0.32	0.15	4.08**
Velars in first	0.12	0.00	10.21***
Dentals in first	0.01	0.12	−3.19*
Vowel density	0.37	0.42	−4.19***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 4
Phonological cues significant for open/closed class words in Dutch

Cue	Dutch		
	Open	Closed	*t*
Length in syllables	1.67	1.31	6.08***
Length in phonemes	4.53	3.28	7.55***
Syllabic complexity	0.63	0.58	3.93***
Voiced consonants in word	0.59	0.70	−2.78*
Nasals in word	0.14	0.24	−2.59[a]
Velars in word	0.18	0.08	4.47***
Nasals in onset	0.07	0.18	−3.29**
Trills in onset	0.05	0.02	2.82*
Alveolars in onset	0.23	0.38	−4.01***
Voiced consonants in onset	0.28	0.48	−5.16***
Plosives in first consonant	0.34	0.19	3.37**
Fricatives in first	0.35	0.19	3.60**
Bilabials in first	0.20	0.07	4.22***
Velars in first	0.17	0.03	6.12***
Vowel density	0.37	0.42	−3.93***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

multiple comparisons within each language, taking correlation between the values into consideration (Perneger, 1998), and the correction for unequal variance was used where appropriate.

As anticipated, several phonological cues distinguished open from closed class words in the English corpus. Table 3 shows the cues with significantly different distributions for open/closed class words. For the word-level cues (length in phonemes, syllabic complexity, vowel density) the values indicate the mean length, the proportion of each syllable that is comprised of consonants in the word, and the (related) proportion of phonemes that are vowels in the word[3]. For the phoneme-level cues, the values indicate the proportion of consonants with the particular manner or place feature. In total, there were 17 cues that significantly distinguished open and closed class words, and one cue that was marginally significant. Results for Dutch, French, and Japanese are shown in Tables 4–6, respectively. As with English, several cues were signicantly differently distributed for open and closed class words in each language.

Fig. 1 presents a Venn diagram of the four languages in terms of the significant phonological cues distinguishing open from closed class words. Cues in italics indicate that closed class words had a higher value for the cue, and cues in plain text indicate that open class words had a higher value. Cues in SMALLCAPS indicate cues that had reverse distributions across the open/closed class distinction for different languages. The general cues about word structure: length in phonemes, length in syllables/morae, syllabic complexity, and vowel density, were significant for three or more of the languages. As anticipated, there was a large overlap between English and Dutch, with 9 cues shared by these languages. English and French overlapped on 3 cues, and English and Japanese on 2 cues. French and Dutch

[3] Rounding resulted in syllabic complexity plus vowel density resulting in a value less than 1.

Table 5
Phonological cues significant for open/closed class words in French

Cue	French		
	Open	Closed	*t*
Length in syllables	1.61	1.21	8.58***
Length in phonemes	3.99	2.64	9.21***
Coronals in word	0.43	0.60	−4.05**
Trills in word	0.15	0.06	4.74***
Bilabials in word	0.20	0.08	4.91***
Alveolars in word	0.36	0.57	−4.66***
Labials in word	0.11	0.05	3.20*
Uvulars in word	0.15	0.06	4.74***
Trills in onset	0.07	0.02	5.12***
Bilabials in onset	0.13	0.05	4.61***
Alveolars in onset	0.21	0.40	−4.35***
Uvulars in onset	0.07	0.02	5.12***
Trills in first	0.04	0.01	2.98[a]
Bilabials in first	0.28	0.12	4.37***
Labials in first	0.11	0.01	7.05***
Alveolars in first	0.23	0.52	−5.75***
Uvulars in first	0.04	0.01	2.98[a]
Vowel height	1.76	1.31	3.77**

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 6
Phonological cues significant for open/closed class words in Japanese

Cue	Japanese		
	Open	Closed	*t*
Length in morae	6.03	4.26	7.84***
Length in phonemes	3.36	2.24	8.72***
Morae complexity	0.48	0.51	−3.89**
Bilabials in word	0.13	0.06	3.62**
Bilabials in onset	0.09	0.03	4.26***
Uvulars in onset	0.05	0.13	−2.80[a]
Glottals in onset	0.04	0.02	2.78[a]
Uvulars in first	0.06	0.18	−2.92[a]
Voiced consonant first	0.39	0.63	−4.40***
Vowel density	0.52	0.49	3.89**
Vowel height	1.23	1.62	−3.57**

[a] $p < .1$.
** $p < .01$.
*** $p < .001$.

overlapped on 5 cues, but perhaps surprisingly there were several cues common to French and Japanese (7 cues, though 3 operated in opposite directions) and Dutch and Japanese (5 cues), even though these languages had distinct genealogy. Only one cue was significant for all four languages: length in phonemes. It is possible that the strict criterion for establishing significance, with correction for multiple comparisons, meant that some significant and useful cues were not identified with the *t*-test analyses, and Experiment 3 returns to

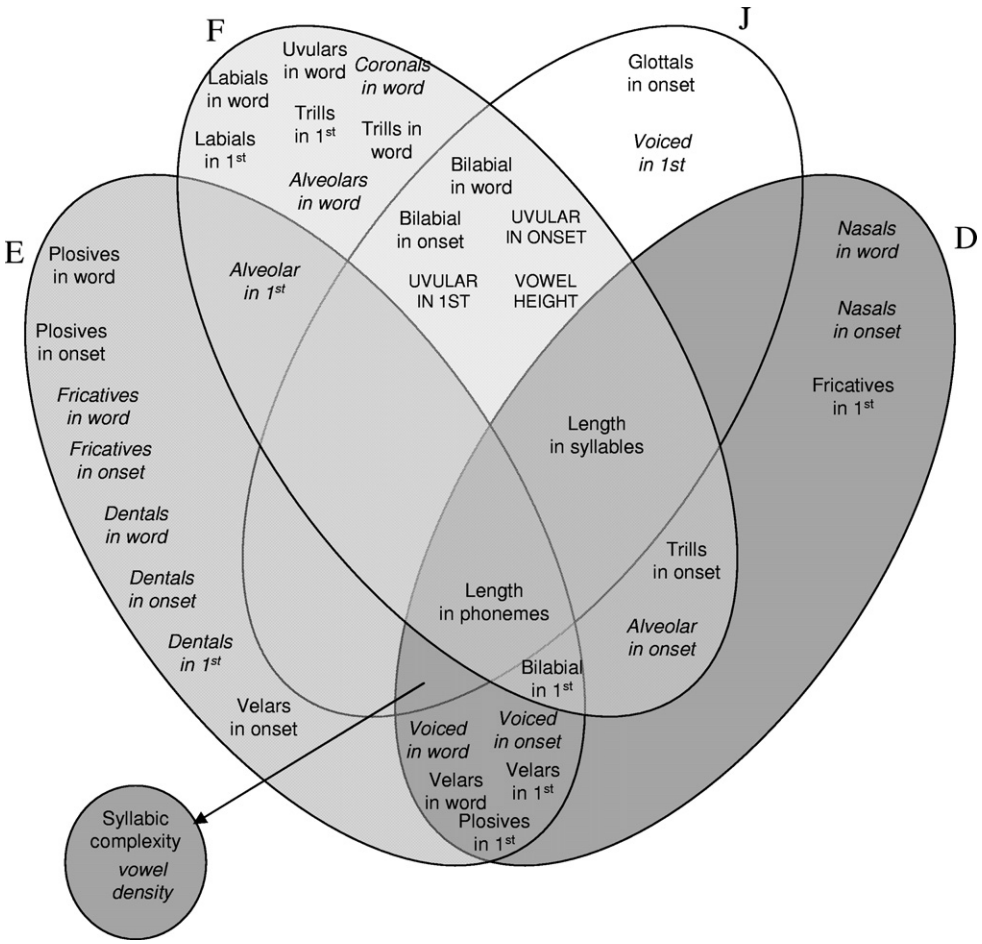

Fig. 1. Venn diagram of phonological cues for distinguishing open from closed class words in English (E), Dutch (D), French (F), and Japanese (J). Cues in plain text indicate that open class words had higher values on the cue than closed class words. Italics indicate cues that had higher values for closed class than open class words. SMALL-CAPS indicates that the distribution of cues was different for the languages.

this issue. Yet, each language reflected a number of phonological cues that related to distinctions in grammatical category, though the precise nature of these cues was language specific.

The corpus of English was larger than the other languages, and though the number of types in each language was identical, the representation of the 1000 highest frequency words in the language is more reliable when derived from a larger corpus. In order to test whether the larger number of significant cues unique to English was due to the larger corpus size we selected the first 358,401 words from the English corpus, making it the same size as the Japanese corpus. 12 of the original 17 cues for English were significant, with 2 of the original cues now marginally significant. Syllabic complexity, marginally significant in the full corpus analysis, was significant in the smaller corpus. The three cues that were no longer significantly differently distributed for open and

closed class words were plosive in the first consonant, bilabials in the first consonant, and velars in the first consonant. The results were stable for these very different corpus sizes, except that significance for the first consonant was not significant: All but two cues unique to English were still significant, except for plosives and velars in first consonant position.

5.2.2. Nouns and verbs

We performed *t*-tests with Bonferroni corrected *p*-values for all phonological cues from Table 1 for all words classified as either nouns or verbs in each language. Results for each of the four languages are shown in Tables 7–10. As for the open/closed class distinction, several phonological cues for each language were significantly differently distributed for nouns and verbs. Fig. 2 shows a Venn diagram of the overlap between different languages in terms of the phonological cues. Overlap between English and Dutch was smaller than for the open/closed class distinction, and English and French overlapped on several cues though sometimes acting in different directions. Japanese and English overlapped on cues that were also shared with other languages. As with the open/closed class distinction, Dutch, French, and Japanese overlapped to a large extent on the cues.

5.3. Discussion

The exhaustive search of phonological cues yielded many cues that distinguished both open from closed class words, and nouns and verbs across all four languages. The results of the English analyses replicated earlier studies that have shown that several cues distinguish grammatical categories in English (Cutler, 1993; Durieux & Gillis, 2001; Kelly, 1996; Monaghan et al., 2005; Sereno & Jongman, 1990). Hence, the search confirmed our hypothesis about the relationship between word class and speech sound, and indicated a valid extension of this approach to languages other than English. The correspondence between phonological cues and grammatical category was found in all four languages. For Dutch and French, several cues were found for distinguishing both open from closed class words, and nouns from verbs. Fewer phonological cues were found for Japanese for the open/closed class distinction, with manner and place feature distinctions being rare, though still some differences were found, in particular, in terms of word length, and vowel height. The morae of Japanese words, composed of a vowel alone, a consonant alone, or a consonant and a vowel, perhaps provide fewer opportunities for distinctions between categories to emerge. The emphasis on the vowel in this language may mean that the distinction in terms of vowel height between open and closed class words is especially important as a cue to category. However, the noun/verb distinction was captured by many phonological cues in Japanese. In English, fewer cues were found to distinguish nouns from verbs than in similar analyses using linguistically-derived phonological cues (Kelly, 1992; Monaghan et al., 2005), and the stringent correction for multiple comparisons may have underestimated the potential contribution of the cues in the current analyses.

The cues that were found to relate to different grammatical categories were sometimes found for languages other than English, but occasionally acting in reverse directions. In all four languages, open class words were longer than closed class words in terms of phonemes, and a significant effect in terms of syllable length was found for Dutch, French, and

Table 7
Means of phonological cues for nouns and verbs for English

Cue	English		
	Noun	Verb	*t*
Length in syllables	1.47	1.27	4.53***
Bilabials in word	0.19	0.12	3.53**
Velars in word	0.14	0.20	−2.99[a]
Approximants in first	0.08	0.17	−3.16*
Bilabials in first	0.27	0.16	3.31*
Vowel density	0.37	0.34	4.25***
Reduced vowels	0.10	0.03	5.40***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 8
Means for phonological cues for nouns and verbs for Dutch

Cue	Dutch		
	Noun	Verb	*t*
Trills in word	0.11	0.07	2.50[a]
Fricatives in word	0.21	0.30	−4.01***
Bilabials in word	0.18	0.12	3.31**
Velars in word	0.16	0.22	−2.69*
Palatals in word	0.06	0.01	6.52***
Labials in word	0.05	0.10	−3.31**
Plosives in onset	0.20	0.15	2.85*
Fricatives in onset	0.09	0.15	−4.01***
Bilabials in onset	0.10	0.05	3.91***
Plosives in first	0.48	0.27	5.31***
Fricatives in first	0.27	0.48	−5.15***
Approximants in first	0.08	0.14	−2.47[a]
Bilabials in first	0.30	0.13	4.95***
Labials in first	0.09	0.18	−2.87*
Voiced consonant first	0.49	0.61	−2.66[a]
Reduced vowels	0.21	0.27	−2.56[a]

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Japanese. Syllabic complexity, a marker of open class words in English and Dutch, was not found in French, and was found in reversed form for Japanese. The density of consonants is greater for open class words in English and Dutch, equal for French, but less for Japanese. For the noun/verb distinction, all the languages demonstrated more bilabials in nouns than verbs, and French and Japanese had more velars in nouns whereas English and Dutch had a tendency for more in verbs.

Length effects were mixed across the languages: nouns tended to have more syllables in English, but verbs contained more phonemes in Japanese. Syllabic complexity was significantly different for French, with more consonants in nouns. There was a trend to

Table 9
Means for phonological cues for nouns and verbs for French

Cue	French		
	Noun	Verb	*t*
Syllabic complexity	0.58	0.54	3.78**
Plosives in word	0.37	0.29	3.07*
Trills in word	0.12	0.18	−3.20*
Bilabials in word	0.22	0.15	3.35*
Velars in word	0.09	0.04	4.16***
Palatals in word	0.12	0.08	2.71[a]
Labials in word	0.09	0.15	−3.17*
Uvulars in word	0.12	0.18	−3.20*
Fricatives in onset	0.14	0.20	−2.91*
Labials in onset	0.05	0.10	−3.23*
Plosives in first	0.42	0.31	3.22*
Trills in first	0.03	0.07	−2.88*
Bilabials in first	0.33	0.19	4.43***
Uvulars in first	0.03	0.07	−2.88*
Vowel density	0.42	0.46	−3.81**
Reduced vowels	0.01	0.04	−3.17*

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 10
Means for phonological cues for nouns and verbs for Japanese

Cue	Japanese		
	Noun	Verb	*t*
Length in phonemes	5.83	6.31	−3.11*
Coronals in word	0.42	0.55	−5.24***
Plosives in word	0.39	0.48	−3.57**
Nasals in word	0.25	0.17	4.00**
Fricatives in word	0.24	0.14	5.90***
Flaps in word	0.07	0.15	−5.21***
Bilabials in word	0.17	0.09	4.75***
Velars in word	0.18	0.13	2.94*
Alveolars in word	0.36	0.50	−6.28***
Glottals in word	0.05	0.08	−3.23*
Bilabials in onset	0.12	0.06	3.48**
Nasals in first	0.10	0.22	−4.21***
Fricatives in first	0.24	0.16	2.81[a]
Approximants in first	0.03	0.09	−3.07*
Uvulars in first	0.03	0.11	−3.76**
Vowel position	0.82	0.66	3.14*
Rounded vowels	0.25	0.15	4.79***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

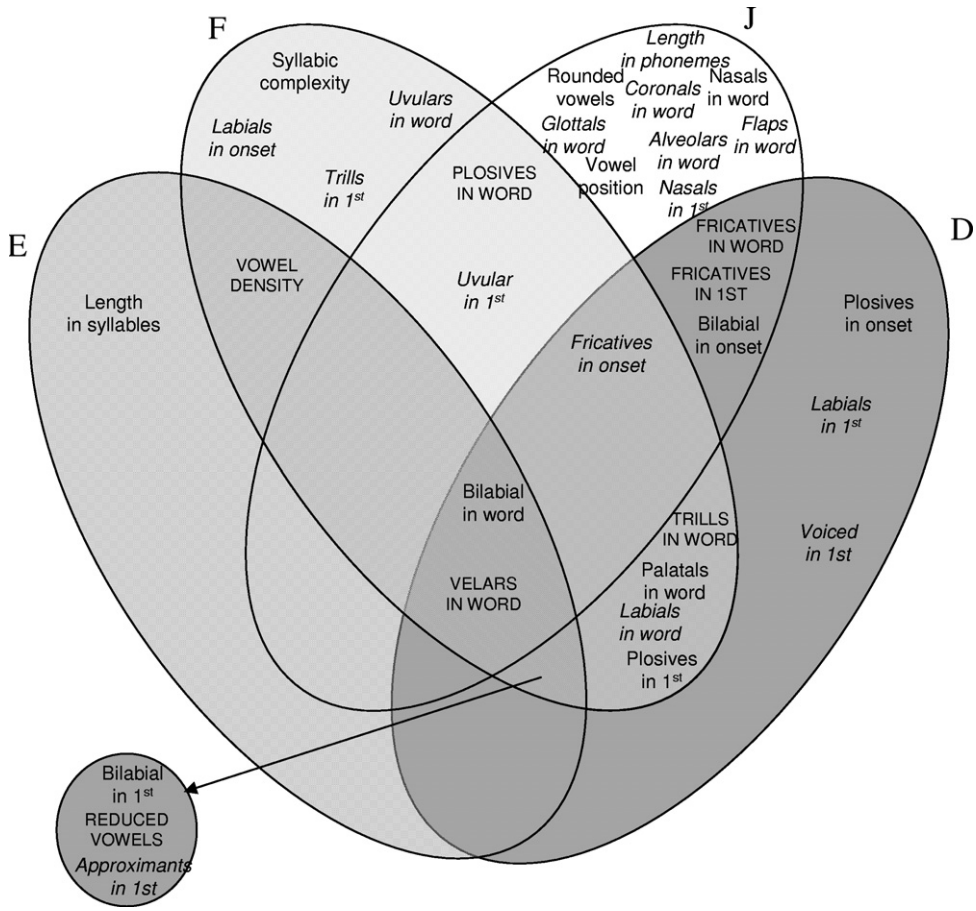

Fig. 2. Venn diagram of phonological cues for distinguishing nouns from verbs in English (E), Dutch (D), French (F), and Japanese (J). Plain text indicates cues with higher values for nouns than verbs, italics indicates cues with higher values for verbs than nouns, and SMALLCAPS indicates cues that were differently distributed for the languages.

a similar effect in Japanese, and a tendency towards the reverse effect in English and Dutch.

The results indicate that phonological cues are, not surprisingly, language specific, with differences in category expressed in different ways in each language. The child learning his or her native language has to learn the correspondences for that particular language. Yet, the range of cues found across different languages to relate to different categories is persuasive evidence that such cues may play a major role in the categorization of words, and this supports the first claim of the PDCH. Different languages have diverse constellations of cues, but all languages have a set of statistically significant cues that reflect broad syntactic categories.

Experiment 2 tested whether the potential importance of distributional cues, found in previous studies of English is also generalizable to other languages.

6. Experiment 2: Cross-linguistic analyses of distributional cues in categorization

6.1. Method

6.1.1. Corpus preparation

The corpora used were identical to those in Experiment 1.

6.1.2. Cue generation

From each language corpus, we generated a total of 50 distributional cues. We selected the 25 most frequent words from each language corpus to be used as the context words. Such high-frequency words may provide stable points within the language around which other words can be categorized, from which the language structure can be derived (Valian & Coulson, 1988). Many of the 25 context words were closed class words, which tend to be phonetically reduced (Morgan et al., 1996). However, closed class words are distinctive from open class words in terms of their phonology which is perceptually available to the child (Shi, Werker, & Morgan, 1999). We then counted the number of times that each of the other words in the corpus occurred immediately preceding or immediately succeeding each of the context words. We used a signed log-likelihood test statistic (Monaghan et al., 2005) to estimate whether the distribution of the target word with the context word was greater or less than chance. This test was adapted from Dunning (1993) and is particularly appropriate for analysing small language corpora. If the association is greater than chance then the signed log-likelihood test produces a positive value, if the words occur together less than would be expected by chance then the test value is negative. If the words co-occur at chance levels, such that the words co-occur with a frequency identical to that which would be expected if words occurred randomly in the corpus, then the test produces a value close to zero.

To illustrate this test, consider the context word *to* in English. We predict that *to* is a useful anchor word for verbs, so it should occur prior to verbs more than would be expected by chance. To assess whether the verb *eat* is significantly associated with this context word, we count the number of times that the phrase *to eat* occurs in the corpus. In the English corpus described above, this pair occurs 2151 times. However, both *to* and *eat* occur highly frequently—108,245 and 6070 times, respectively, in the corpus—and so, just by chance, they may occur together several times. The signed-log-likelihood test assesses whether the 2151 co-occurrences are more than would be expected if the distribution of words was just random. The signed log-likelihood value for *to eat* is 9036.6, indicating that this association is highly significant, and therefore that *to* is likely to occur prior to *eat*. For the phrase *eat to*, however, the association is not as strong. This phrase occurs 72 times in the corpus, and the signed log-likelihood test value is −209.2, suggesting that *eat* occurs before *to* less than would be expected by chance.

Once all 50 context cues had been generated for the 1000 most frequent words in each language we tested whether open and closed class words and nouns and verbs had different mean signed log-likelihood test values for each context word cue. We anticipated, for example, that *to* would act as a good preceding cue for verbs: We hypothesised that verbs are more likely than chance to occur after this word, whereas nouns are less likely than chance to occur in this position. The 25 highest frequency words used as cues for each language are listed in Appendix A.

Table 11
Signed log-likelihood test values for distributional cues that significantly distinguished open from closed class words in English

Context word cue	English		
	Open	Closed	*t*
he___	59.78	−156.57	4.56***
we___	81.10	−189.23	3.60*
your___	141.52	−191.87	4.23***
a___	287.38	−583.87	4.41***
___that's	−26.26	−207.51	3.52*
___oh	−52.02	−347.66	4.05**
___and	1.75	−244.67	3.88**

Word ___ indicates the context word cue is the preceding word, ___ word indicates that the succeeding word is the context word cue.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 12
Signed log-likelihood test values for distributional cues that significantly distinguished open from closed class words in Dutch

	Dutch		
	Open	Closed	*t*
ja___	−35.42	−186.27	4.20***
de___	78.70	−129.09	5.71***
het___	39.06	−74.44	5.43***
die___	17.37	−43.32	5.38***
nou___	−1.73	−22.02	3.76**
zo___	1.55	−43.88	4.10**
dan___	11.85	−34.70	4.04***
oh___	−8.47	−45.12	4.60***
nee___	−8.59	−44.67	4.00**
he___	−8.33	−51.08	4.28***
___ja	−35.56	−241.58	4.34***
___en	−6.82	−73.21	4.30***
___oh	−11.18	−68.71	4.44***
___nee	−10.29	−62.61	4.29***
___he	−2.19	−43.08	4.35***

** $p < .01$.
*** $p < .001$.

6.2. Results

6.2.1. Open and closed class words

Distributional cues that distinguished open from closed class words in English, Dutch, French, and Japanese are shown in Tables 11–14, respectively. For English, the preceding words *he* and *we*, both pronouns, are likely to occur before verbs and adverbs but less

Table 13
Signed log-likelihood test values for distributional cues that significantly distinguished open from closed class words in French

	French		
	Open	Closed	*t*
la___	20.18	−39.85	4.51***
ça___	−3.48	−43.70	6.13***
le___	16.48	−36.48	4.53***
l'___	19.61	−36.49	3.22*
les___	13.74	−24.72	5.00***
n'___	10.22	−21.19	3.10*
des___	9.03	−16.47	4.53***
___pas	19.52	−59.72	4.46***
___ça	−4.34	−49.11	6.33***
___oui	−3.62	−45.48	6.27***
___un	4.81	−24.96	5.54***
___et	−1.03	−28.44	6.42***
___de	6.02	−19.11	4.86***
___non	−1.01	−23.55	3.29*
___dans	2.59	−15.58	6.33***

* $p < .05$.
*** $p < .001$.

Table 14
Signed log-likelihood test values for distributional cues that significantly distinguished open from closed class words in Japanese

	Japanese		
	Open	Closed	*t*
mo___	4.94	−14.23	5.31***
tte___	5.29	−13.92	3.36*
ga___	10.16	−22.97	3.27*
a___	3.06	−18.52	3.32*
ni___	11.13	−15.84	3.08*
hai___	−0.64	−29.33	4.22*
yo___	−3.12	−30.25	3.02[a]
ne___	−3.08	−36.92	3.62**
un___	−3.89	−63.01	3.94**
___soo	−1.37	−10.49	3.92**
___ja	−1.15	−7.71	4.08**
___n	−1.56	−24.20	4.17**
___a	−3.04	−28.67	4.45***
___hai	−3.61	−33.42	4.56***
___un	−7.89	−75.88	4.52***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

likely to occur before closed class words. *A* is more likely to precede nouns and adjectives than closed class words. All 3 succeeding word cues were strongly dissociated with closed class words, and either weakly dissociated or weakly associated with open class words.

That's and *and* function as connectives, and the interjection *oh* is more likely to occur at the end of a phrase, and so all are much less likely to occur after closed class words.

For Dutch, 10 preceding word cues and 5 succeeding cues significantly distinguished open from closed class words (Table 12). The articles *de* (the), *het* (the), and *die* (this) occurred significantly more often before open class words, due to their association with nouns. Similarly to English, the interjections *ja* (yes), *nou* (now), *oh* (oh), *nee* (no), and *hé* (hey) were more likely to precede open class than closed class words. The connectives *zo* (so) and *dan* (then) were also more likely to precede open class words. For the succeeding cues, closed class words were more likely to occur after the interjections *ja, oh*, *nee*, and *hé*, and the connective *en* (and).

In French, 7 preceding and 8 succeeding cues were significant (Table 13). Many of the preceding cues were articles. For the succeeding cues, *pas* (negative particle) and *un* (a) are likely to occur after verbs but unlikely to occur after closed class words. Other succeeding word cues are connectives, demonstrating a similar pattern to English.

For Japanese, the preceding cues *mo* (indicating addition, gloss = as well), *ga* (nominative case marker[4]), and *ni* (dative case marker) are particles that modify the preceding noun. Though these particles modify the *preceding* noun, they are strongly associated with an open class word following them. *Yo* (indicating command) and *ne* (indicating request for confirmation, "isn't it") are particles that occur at the end of sentences, and are consequently likely to occur before the first content word of another sentence (the topic of the sentence often occurs at the beginning of the sentence). *Tte* can occur sentence finally or after a noun to indicate a quotation, and *a* (ah), *hai* (yes) and *un* (yeah) are interjections and tend to occur between phrases and so are likely to precede nouns. For the succeeding cues, all the context words are interjections, which are likely to occur between phrases, reflecting their strong dissociation with closed class words, and weak dissociation with open class words.

6.2.2. Nouns and verbs

The results for the four languages are presented in Tables 15–18. For English, 3 preceding cues were associated with verbs more closely than with nouns: the pronouns *he* and *we*, and the infinitive verb marker *to*. 9 preceding cues were more strongly associated with nouns: the articles *a* and *the*, the possessive pronouns *your* and *that's*, *no* (perhaps functioning as a quantifier), the verbs *are*, *do*, and *is*, and the preposition *in*. For the succeeding cues, *your*, *a*, *it*, and *the* were significantly more associated with verbs than nouns. Such words are likely to begin noun phrases as verb objects. *Are*, *do*, *is*, *no*, and *oh* were significantly more associated with nouns than verbs, which was largely due to such words being more unlikely to occur following verbs, except for *is*, for which there was a significant association with nouns (Table 15).

For Dutch, 12 preceding cues and 14 succeeding cues were significantly differently associated with nouns and verbs (Table 16). The articles *een* (a) and *de* (the) are strongly associated with a following noun and negatively associated with verbs. The pronoun *ik* (I) is strongly associated with a succeeding verb and negatively associated with nouns, as are the connectives *nou* (now) and *dan* (then). The connective *nog* (yet) is strongly dissociated with

[4] The particles in Japanese have many uses. *Ga*, for instance, can also be used as an accusative case marker for some verbs, and *ni* can indicate the passive, a time, or a location. In these cases, however, the particle occurs in the same distributional relationship with the noun.

Table 15
Signed log-likelihood test values for distributional cues that significantly distinguished nouns from verbs in English

Context word cue	English		
	Nouns	Verbs	*t*
He___	−13.72	252.54	−4.95***
we___	−15.75	339.99	−4.04**
are___	−15.13	−46.82	3.77**
no___	−2.22	−44.82	5.52***
your___	304.50	−50.83	7.25***
that's___	−14.21	−56.01	3.88**
in___	−7.26	−73.79	5.21***
do___	−21.13	−69.94	5.58***
is___	−20.74	−64.49	3.43*
to___	33.45	779.79	−4.19**
a___	392.41	−111.74	6.14***
the___	730.99	−207.07	11.02***
you___	−87.52	1522.01	−3.41*
___are	2.02	−53.07	5.90***
___no	−17.97	−41.09	3.29*
___your	−10.39	128.17	−3.68*
___oh	−27.83	−81.55	4.30***
___do	−7.84	−58.58	3.77**
___is	19.48	−71.82	6.39***
___a	−38.33	367.91	−3.30*
___it	−39.59	698.74	−3.42*
___the	−58.24	130.94	−3.13[a]

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

a following verb than a noun. The interjections *ja* (yes), *oh* (oh), *nee* (no), and *hé* (hey) were more likely to occur before nouns than verbs, though they were negatively associated with both nouns and verbs. The verbs *is* (is) and *moet* (must) were more strongly dissociated with a following verb than a following noun. For the succeeding cues, the pronouns *je* (you) and *ik* (I) are more likely to succeed verbs. Succeeding interjections *ja* (yes), *nou* (now), *oh* (oh), *nee* (no), and *hé* (hey) were dissociated with both nouns and verbs but more so with verbs. The verbs *is* (is) and *moet* (must) are more strongly dissociated with preceding verbs than nouns, and the article *het* (the) is more likely to occur after verbs than nouns. *Wat* (what), classified as a pronoun, is less likely to succeed a verb, and the connectives *en* (and), *zo* (so), and *in* (in) are more likely to succeed nouns.

For French, articles were more strongly associated as preceding nouns than verbs: *la*, *le*, *l'*, *les* (the), *un* (a), and *des* (some). The connectives *de* (of/from), *qui* (which), *que* (that), and *et* (and) were also more likely to precede nouns than verbs. Verbs were more strongly associated with preceding pronouns *tu* (you) and *on* (pronoun one), and also the negative particle *n'*. For the succeeding word cues, the pattern was almost reversed, with articles more likely to succeed verbs than nouns—noun phrases beginning with an article are likely to follow a verb: *la*, *le*, *un*, *l'*, *les*, and *des*. The negative particle *pas* and the preposition *dans* (in) was also more likely to precede verbs. The verb *est* (is), conjunctions *et* (and) and *qui*

Table 16
Signed log-likelihood test values for distributional cues that significantly distinguished nouns from verbs in Dutch

	Dutch		
	Nouns	Verbs	*t*
ja___	−14.61	−48.86	3.77**
is___	−6.42	−32.66	4.04**
een___	124.53	−28.96	7.40***
de___	223.03	−29.91	9.17***
ik___	−3.11	121.12	−3.46*
nou___	−4.28	1.04	−3.77**
dan___	−3.63	48.07	−4.23***
oh___	−3.94	−10.63	3.70**
nee___	−3.68	−11.79	4.01**
he___	−2.19	−11.72	4.12***
nog___	−1.91	−7.60	3.31*
moet___	−1.00	−7.82	2.98*
___ja	−11.66	−50.70	3.84**
___je	−10.48	395.35	−3.47**
___is	−0.25	−33.47	5.15***
___het	−6.06	53.40	−2.84[a]
___wat	−5.49	−15.22	3.66*
___ik	−0.01	116.81	−2.79[a]
___en	2.43	−13.23	3.87*
___nou	−3.08	−9.30	4.62***
___zo	−3.29	−6.11	3.29*
___oh	−4.37	−14.54	3.91**
___nee	−4.14	−13.83	3.77**
___in	8.30	−2.73	3.97**
___he	−0.51	−6.59	3.43*
___moet	−1.08	−10.25	4.55***

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

(which), interjections *oui* (yes), and *non* (no), the article *c'* (this/that), the pronoun *il* (it), the preposition *de* (of/from), and the negative particle *n'* were more likely to succeed nouns than verbs (Table 17).

In Japanese, 29 distributional cues were also found to be significantly differently associated with nouns and verbs (Table 18). More strongly associated with a following noun were the proper name *Taro* (a child in one of the Japanese corpora), the pronoun *sore* (it/that), the sentence-final particle *na* (negative imperative), the particles *wa* (topicaliser) and *no* (genitive marker) that modify the preceding noun, the interjections *ja* (also used colloquially as noun marking particle), *hai* (yes), and *un* (yeah), the sentence-final particles *ne* (request for confirmation), and the clause-final particles *ka* (indicating question, also used as "or") and *yo* (indicating command). Words that were more strongly associated with a following verb were the noun modifying particles *mo* (indicating addition), *ga* (nominative case marker), and *ni* (dative case marker), and the sentence-final particle *tte* (indicating quotation). For succeeding word cues, the pronouns *doko* (where), *koko* (here), and *nani* (what) were more strongly associated with a preceding noun. Predictably, the noun

Table 17
Signed log-likelihood test values for distributional cues that significantly distinguished nouns from verbs in French

	French		
	Nouns	Verbs	*t*
tu___	−4.15	106.30	−3.15*
la___	46.63	−4.28	5.40***
le___	29.08	−0.73	6.43***
un___	15.29	−4.34	6.69***
l'___	37.24	4.18	2.64[a]
que___	−0.95	−5.14	2.82*
on___	−1.76	35.46	−2.96*
les___	24.53	−0.31	6.62***
et___	−0.63	−3.67	2.78[a]
de___	5.37	−2.20	4.46***
n'___	−1.34	30.29	−3.46**
qui___	−1.06	12.59	−2.64[a]
des___	17.45	−2.93	6.81***
___est	−5.12	−20.08	2.77[a]
___c'	−1.51	−5.96	4.43***
___pas	−2.56	56.66	−3.72**
___la	−1.12	5.10	−5.27***
___le	−1.46	5.29	−5.26***
___oui	−0.52	−6.11	3.00*
___il	−0.65	−4.02	3.60**
___un	−1.64	13.41	−3.91**
___l'	−1.34	2.78	−2.81[a]
___les	−0.85	6.38	−5.41***
___et	1.87	−3.70	4.71***
___de	9.00	2.07	3.64**
___non	−1.02	−4.07	2.73[a]
___n'	0.04	−4.00	2.79[a]
___qui	3.46	−1.87	6.19***
___des	−0.38	8.40	−3.22*
___dans	1.50	5.05	−2.76[a]

[a] $p < .1$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

modifying particles that succeed the noun were more strongly associated with nouns: *de* (indicating manner), *mo* (indicating addition), *ga* (nominative case marker), *ni* (dative case marker), and *wa* (topicaliser). Also more likely to succeed nouns were the pronoun *kore* (this), and the copula *da*. More strongly associated with a preceding verb were the end of phrase/sentence particles *tte* (indicating quotation), *ne* (requesting confirmation) and *yo* (indicating command), and the interjection *n* (yeah).

6.3. Discussion

Just as with the phonological analyses we reported in Experiment 1, the distributional analyses we have undertaken generalize across different languages. In each language we found a range of distributional cues that distinguished open from closed class words, and

Table 18
Signed log-likelihood test values for distributional cues that significantly distinguished nouns from verbs in Japanese

	Japanese		
	Nouns	Verbs	*t*
taro___	0.63	−0.71	3.73**
sore___	0.26	−0.55	3.30*
na___	−0.73	−1.60	5.12***
ja___	5.34	−0.26	3.20*
mo___	0.77	10.45	−3.38*
tte___	−0.92	17.05	−3.43*
ga___	1.15	17.61	−3.58**
ka___	−1.85	−3.17	3.58**
ni___	−0.97	33.47	−4.51***
hai___	−0.70	−2.64	3.96**
wa___	0.38	−2.40	5.64***
yo___	−1.85	−4.79	6.02***
ne___	−1.52	−5.28	6.99***
un___	−2.26	−6.58	5.78***
no___	15.24	−5.29	5.14***
___doko	1.96	−1.26	6.79***
___de	11.73	−1.62	3.81**
___koko	−0.54	−1.28	4.80***
___mo	8.41	−1.26	5.97***
___nani	−0.64	−2.25	3.57**
___tte	−0.14	2.68	−3.17*
___da	11.73	−2.58	4.59***
___n	−2.22	−0.30	−4.02**
___ga	11.15	−3.30	12.22***
___ni	17.72	−2.69	5.37***
___wa	10.51	−1.78	4.70***
___yo	−1.93	36.68	−5.09***
___kore	−2.46	−4.03	2.99*
___ne	−1.47	19.73	−4.57***

* $p < .05$.
** $p < .01$.
*** $p < .001$.

nouns from verbs. In all four languages, the high-frequency closed class words—articles and pronouns in English and French, particles in Japanese—reflect large differences in distributions between open and closed class words. Interjections also figured in distinguishing these categories of word as succeeding cues, indicating that they are more likely to occur after open than closed class words. Across all languages, there appear to be constraints as to where such interjections can occur which tend to be inter-phrasal. Such phrases in all three languages we have analysed tend to finish with open class words, resulting in the difference in strength of association with such interjection words.

Many more cues significantly distinguished nouns from verbs than open from closed class words across all the languages, perhaps reflecting the category-specificity of particular distributional cues. The open/closed class distinction required words from several different categories to be grouped together on the basis of the cues. All languages demonstrated several distributional cues marking both the preceding and the succeeding word as strong

indicators of word class. In Japanese, as would be anticipated from the language structure, the noun modifying particles were strongly associated with preceding nouns compared to verbs, but in English, Dutch, and French the succeeding word cues were also found to distinguish grammatical categories—interjections were often used between phrases, which more frequently ended with nouns than verbs. Articles were strongly associated with a following noun in all four languages (though usage was rarer in Japanese), and pronouns occurred with stronger association before verbs than nouns in English, Dutch, and French. In Dutch and Japanese, pronouns were more likely to occur after nouns than after verbs, a pattern that can be anticipated from the verb-second structure of Dutch and the subject-object-verb structure of Japanese, where two noun phrases may co-occur.

The two experiments we have reported have indicated that generalizable analyses of phonological and distributional cues can be performed on a set of languages with very different characteristics, but that qualitatively similar effects can emerge from these languages. The comparison of means for each cue that we have conducted indicates whether information about grammatical category is reflected in differences in the distribution of each of the phonological and distributional cues. However, such tests do not provide detailed information about how successful classification of different syntactic categories may be on the basis of these cues. In Monaghan et al. (2005), we distinguished analyses of cues that are significantly different (i.e., the distribution of cues is distinct in each language) from those that reflect their diagnosticity (i.e., how useful the cues are for determining the categories). Significantly different distributions may still have a large overlap and so may not be effective or useable for determining category. The next experiment compared the extent to which phonological cues and distributional cues can distinguish open from closed class words, and nouns from verbs in a diagnostic test, using discriminant analyses.

We had the additional aim of determining whether cues that are significantly differently distributed across different categories overlap or are additive in their contribution to successful classification. We predicted that combined cues would result in better classification of grammatical categories than using just one cue type alone. Yet, it was unclear the extent to which the different cues overlapped in classifying words, and so we wanted to determine the extent to which the cues operated in concert for categorisation. However, the internal and external cues may classify the same words, hence adding to the saliency of category for these words, or alternatively, each cue type may operate on a different subset of words, so that in combination the cues successfully categorize a larger set of words. In the latter case, the cues provide a balance in the effectiveness of their classification. We tested this by assessing whether phonological and distributional cues used separately to classify words provided better performance than when the cues were entered jointly in a discriminant analysis. The PDCH claims that phonological coherence will be more effective where other cues, such as distributional sources of information, are weaker for determining categories. We predicted that the discrimination based on phonological cues would classify a different set of words to that of the distributional cues.

7. Experiment 3: Cross-linguistic analyses of phonological and distributional cues

7.1. Method

7.1.1. Corpus preparation and cue generation

The corpora were identical to those employed in Experiment 1. Phonological and distributional cues were generated in the same way as for Experiments 1 and 2.

7.1.2. Cue assessment

The extent to which cues could be used to predict syntactic category was determined by using stepwise discriminant analysis in order to predict category membership of open and closed class words, and nouns and verbs. Stepwise discriminant analysis enters those cues that contribute significantly toward the category distinction in order to maximise correct classification. The analyses were weighted according to their frequency per million so that the power was similar across all languages and the analyses indicated the total number of word tokens that could be correctly classified by each language. We used a leave-one-out cross-validation procedure, where the discriminant analysis was performed on the basis of all words except one, then the classification of the omitted word is determined. This process was then repeated omitting each word in the corpus in turn, and the classifications were tallied. We performed three analyses: one entering only the phonological cues for each language, one with just the distributional cues, and the third with both phonological and distributional cues. A random baseline for classification was determined by randomising the assignment of categories to words but maintaining the original category sizes in terms of type and token frequencies. For each randomised assignment, the discriminant analyses were then performed. This was repeated ten times and the results were averaged.

We then assessed whether the phonological cues and the distributional cues were classifying a complementary set of words by computing a three-way hierarchical loglinear analysis, with category (open/closed or noun/verb), classification based on phonological cues, and classification based on distributional cues as factors. If the analysis indicates that the three-way effect is needed to account for the data then this means the phonological and distributional cues operate on significantly distinct sets of words.

7.2. Results

7.2.1. Open and closed class words

The accuracy of classification of open and closed class words in English, Dutch, French, and Japanese is shown in Fig. 3. Each bar indicates the accuracy of classification, averaged across both open- and closed-class categorisations. The lower-part of each bar indicates the chance classification, and the shaded part indicates the improvement over chance. The effect size and significance of the classification based on each group of cues was measured using Wilk's λ, which varies between 0 and 1 with accurate classification reflected by values close to 0. For English, for phonology, $\lambda = .591$, for distributional cues, $\lambda = .267$, and for combined cues, $\lambda = .179$, all $p < .0001$. For Dutch, for phonology, $\lambda = .421$, for distributional cues, $\lambda = .243$, and for combined cues, $\lambda = .162$, all $p < .0001$. Similarly, for French, for phonology, $\lambda = .417$, for distributional cues, $\lambda = .216$, and for combined cues, $\lambda = .141$, all $p < .0001$. Finally, Japanese demonstrated a similar pattern, for phonology, $\lambda = .291$, for distributional cues, $\lambda = .190$, and for combined cues, $\lambda = .144$, all $p < .0001$. When frequency was not taken into account, so the analyses were word types and not word tokens, the results were very similar, with significantly better performance than chance in all languages for all cue types (all $p < .0001$).

In each language, the combined cues resulted in a better classification than either cue type alone. The hierarchical log-linear analyses indicated that the three-way effects were necessary to account for the classifications of open- and closed-class words. If the phonological cues and distributional cues were classifying the same set of words, then one of these classifications would have reduced to the other, meaning the three-way classification

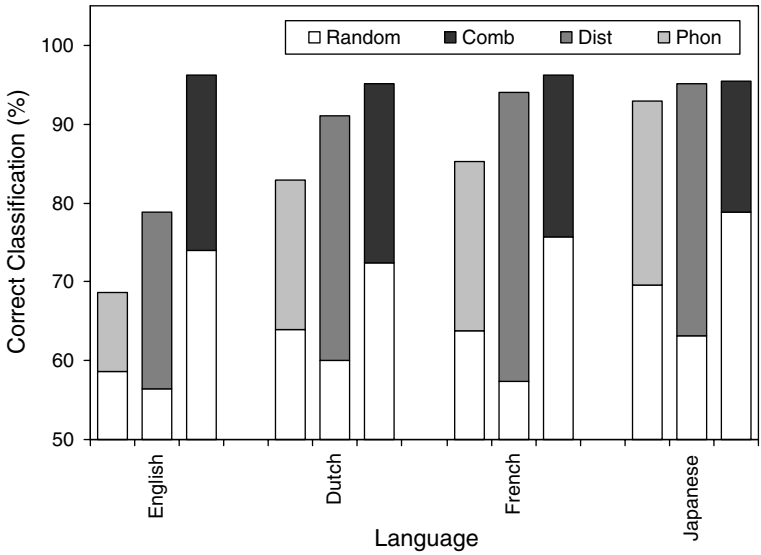

Fig. 3. Accuracy of classifications for open/closed class words, for phonological, distributional, and combined cues. Italics indicate cues with different distributions for the languages.

was not necessary. If the cues were operating complementarily, such that they classified significantly distinct sets of words, then the three-way effect should be necessary for matching the classifications of open and closed class words for each cue type. Removing the three-way effect resulted in a classification significantly poorer than chance in each language. For English, logistic regression $\chi^2(1) = 22{,}458.683$, for Dutch, logistic regression $\chi^2(1) = 17{,}893.543$, for French, $\chi^2(1) = 21487.895$, and for Japanese, $\chi^2(1) = 71{,}175.646$, all $p < .0001$, meaning that the cues were classifying significantly different sets of words. Fig. 4 shows the open/closed classifications based on phonological cues or distributional cues. For all four languages, distributional cues were better for classifying open class words than were the phonological cues. For Japanese, the phonological cues were better than the distributional cues for classifying closed class words, and the classification was similar for both word classes in the other three languages, though there is an interaction between the effectiveness of cue types, reflected in the three-way hierarchical loglinear analyses: Phonological cues are effective when distributional cues are weaker.

7.2.2. Nouns and verbs

Accuracy of classifications for nouns and verbs are shown in Fig. 5. As with the open/closed class analyses, the classification above chance based on each set of cues was highly significant. For English, for phonology, $\lambda = .659$, for distributional cues, $\lambda = .292$, and for combined cues, $\lambda = .231$, all $p < .0001$. For Dutch, for phonology, $\lambda = .566$, for distributional cues, $\lambda = .389$, and for combined cues, $\lambda = .302$, all $p < .0001$. For French, for phonology, $\lambda = .528$, for distributional cues, $\lambda = .386$, and for combined cues, $\lambda = .314$, all $p < .0001$. In Japanese, for phonology, $\lambda = .548$, for distributional cues, $\lambda = .390$, and for combined cues, $\lambda = .262$, all $p < .0001$. Once again, when frequency was not taken into account in a word-type as compared to a word-token analysis, the results were still significantly better than chance in all languages for all cue types (all $p < .001$).

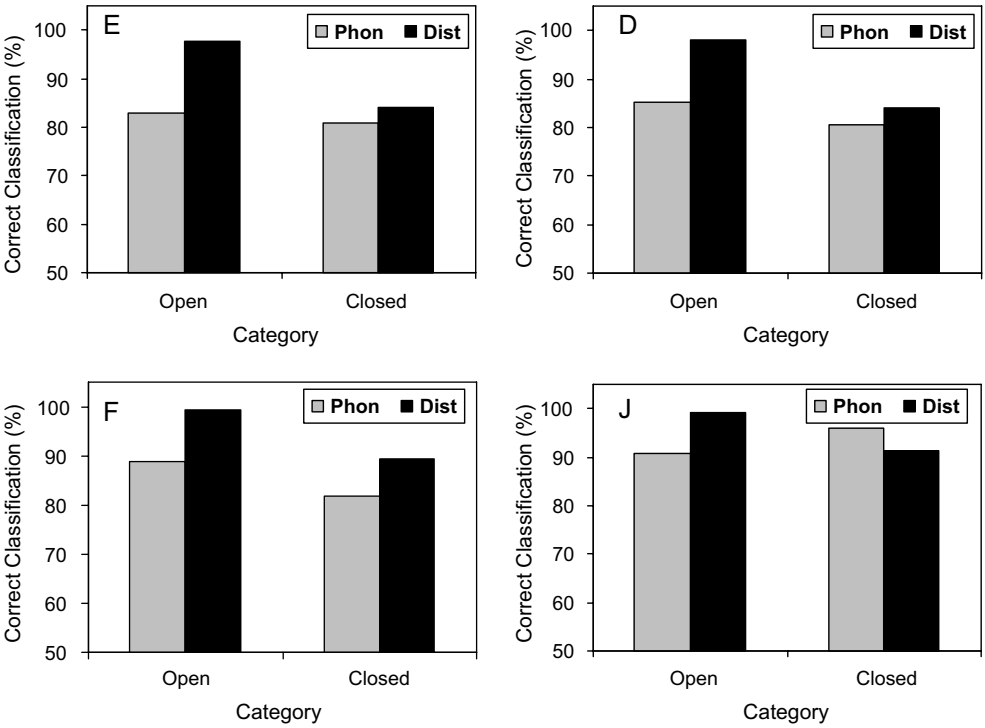

Fig. 4. Classifications of open and closed class words by phonological cues and distributional cues for the four languages (E, English; D, Dutch; F, French; J, Japanese).

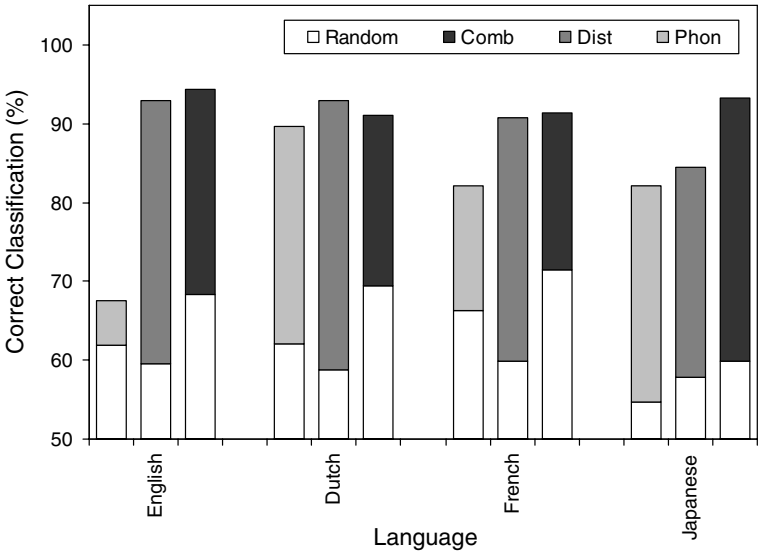

Fig. 5. Accuracy of classifications for nouns/verbs, for phonological, distributional, and combined cues.

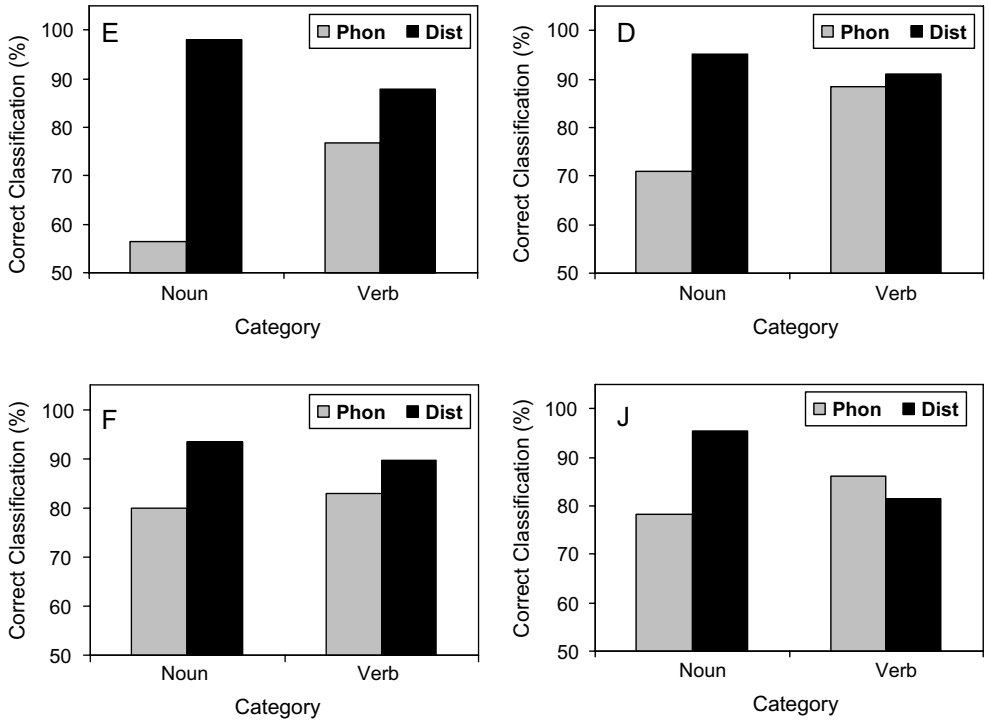

Fig. 6. Classifications of nouns and verbs by phonologivcal cues and distributional cues for the four languages (E, English; D, Dutch; F, French; J, Japanese).

The combined cues resulted in a better classification than either cue type when considered alone except for Dutch analysis where combined cues resulted in slightly poorer classification than just distributional cues (91.0% compared to 93.1%). However, hierarchical log-linear analyses indicated that the three-way effects were necessary to account for the classifications of nouns and verbs. Omitting the three-way effect from modelling each language resulted in a significantly poorer reflection of the actual classification. For English, linear regression $\chi^2(1) = 742.397$, for Dutch, $\chi^2(1) = 644.552$, for French, $\chi^2(1) = 14{,}283.670$, and for Japanese, $\chi^2(1) = 5{,}179.565$, all $p < .001$. Fig. 6 shows that the hierarchical loglinear analysis reflected the fact that distributional cues were better than phonological cues for classifying nouns in all four languages, but that phonological cues were more effective than distributional cues for classifying verbs. In all four languages, the distributional cues were better for classifying nouns than verbs, whereas the phonological cues, weak for nouns, were more reliable for verbs. For Dutch, even though the combined cues resulted in slightly poorer classification, the cue types when considered alone were still found to classify a complementary set of nouns and verbs.

7.3. Discussion

The large set of phonological and distributional cues that were differently distributed for different word categories in Experiments 1 and 2 were found to provide valuable

information for accurate categorization of words. Accuracy was 20-30% above chance levels in most of the analyses, regardless of whether phonological and distributional cues were considered alone or in union, providing support for the PDCH that anticipated a benefit for classification from all these cue types. As anticipated, for all languages and in all but one analysis (excepting Dutch noun/verb classification) performance was better for the combined cues (Figs. 3 and 5), and this improvement was shown to be due to words misclassified by one type of cue tending to be corrected by the other (Figs. 4 and 6). In all four languages, we found that the cues we had located were highly significant and effective indicators of word category.

The higher random baseline values from combined cues also reflects the complementarity of the information present in these cue types. This effect of better performance even in the random case is exactly as predicted by a set of cues that focus on providing information about the category of overlapping but non-identical sets of words. If one cue type was largely redundant then the combined cues would result in similar performance in both randomised and actual classifications. This was not the case. This provides support for the PDCH claim that cues are in a complementary distribution to aid classification. For words where distributional cues are not so rich then phonological cues provide information about category, similarly for words where the phonological cues are indistinct then distributional information supports accurate classification. Based on the results of the classifications, with a high-degree of overlap in correct classifications but almost no overlap in incorrect classifications, a strategy where the category is provided by combined cues therefore appears to be optimal.

In Christiansen and Monaghan (2006), we reported that nouns were better classified by distributional information whereas verbs were better classified by phonological information. A similar result was found in these analyses. Fig. 6 indicates that, for English, a similar pattern pertains: 58.6% of the nouns were correctly classified by the phonological cues and 98.0% were correctly classified by the distributional cues. For the verbs, the classifications based on the phonological cues were more comparable to those of the distributional cues: 82.3 and 88.2%, respectively. The difference to the results of Christiansen and Monaghan (2006) was due to the current analysis of the 1000 most frequent words compared to the larger frequency range of the 5000 words in Christiansen and Monaghan (2006), hence including lower-frequency words where phonological cues are more reliable indicators of category (Monaghan et al., 2005). The better performance for distributional cues than phonological cues for nouns and similar performance for verbs is also shown in Dutch, and the Christiansen and Monaghan (2006) pattern of better performance for phonological cues for verbs is supported in French and Japanese. Verbs tend to occur in a greater range of contexts and hence distributional information is more likely to benefit from supplementary word-internal information about syntactic category, such as phonological information. This pattern has been shown to be general across the four languages we have examined here.

An additional finding from the current study was that phonological and distributional cues also operate differently for open and closed class word classifications. Open class words tend to occur in more constrained contexts, often co-occurring with frequent closed class words (Valian & Coulson, 1988). In contrast, closed class words may occur in very many contexts, and so phonological information is likely to be useful for these words. Shi et al. (1998) have proposed that closed class words share many phonological properties as a result of their reduction in the speech signal. An additional effect of this reduction is one

of phonological similarity effecting better classification performance. Fig. 4 indicates that distributional cues are much better than phonological cues for classifying open class words in all four languages, whereas phonological cues are similarly effective (and more effective in Dutch, French, and Japanese) for classifying closed class words.

8. Experiment 4: Part-of-speech and cues to category

One simplification in the analyses presented thus far is the assignment of a single grammatical category to each word. This simplifying assumption has been made in a number of studies of grammatical categorization and can be justified as a reasonable first approximation, given that the frequency of most syntactically ambiguous words is strongly skewed in favor of one grammatical category (Monaghan et al., 2005; Redington et al., 1998). Thus, in the analyses so far, one may expect that the effect of assigning the single most frequent category to each word in a corpus is to increase the noise in the database that is being assessed: the analyses provide lower-bounds on the potential information present for categorization. However, particularly when investigating interacting cues, this simplification may not hold. In particular, the child has ultimately to learn not only the grammatical category of a word but also must know which words have multiple usages, and which usage is relevant in the current context.

For ambiguous words, therefore, phonological information may be of little use because the phonological form of the word is relatively stable for different usages (though there may be prosodic and phonetic distinctions; e.g., Sereno & Jongman, 1995). In these cases, the child has to derive information about the category of the token from other contextual information. This experiment tested the extent to which the assignment of most frequent grammatical category to each word type affected the results. We report discriminant analyses based on cues derived from a part-of-speech tagged version of the CHILDES corpus for English that distinguishes particular usages for each word token.

8.1. Method

8.1.1. Corpus preparation and cue generation

The corpus was derived from the CHILDES database of child-directed speech (MacWhinney, 2000). The corpus was labelled with the grammatical category of each word with a part-of-speech (POS) tagger MOR, based on MORPH (Hausser, 1989), with disambiguation performed by the programs POSTTRAIN and TRNFIX, both available in the CHILDES release.[5] The POS-tagged corpus was larger than the corpus we originally assessed, as the CHILDES database has had additional corpora added since our first analyses. The POS-tagged corpus omitted false-starts and interruptions, and did not include alternative spellings such as "gonna" for "going to" as did the original corpus, and so distributional information was more likely to be effective for this corpus. The corpus comprised of 5,753,660 words in 1,300,623 utterances, and tagging accuracy is approximately 95% (Sagae, MacWhinney, & Lavie, 2004).

The 1000 most frequent words were selected from the corpus for analysis. For each of these words, all the different usages were recorded along with their frequency of

[5] We are grateful to Brian MacWhinney for making the POS-tagged corpus available within the general release of the CHILDES corpus.

occurrence. So, for example, the word "like" occurred 24,252 times as a verb, 6,698 times as a preposition, 75 times as a conjunction, and 406 times as an interjection. Each of these usages was treated as a separate entry in the database. The phonological form of each word was derived from the CELEX database, and all words that were not present in CELEX were hand-coded, as in Experiment 1.

The phonological and distributional cues were derived in the same way as for Experiments 1 and 2. For the words with multiple usages, the phonological cues were therefore identical, but the distributional cues differed. For the above example, the co-occurrence of each context word with the word "like" when used as a verb was counted separately from co-occurrence of the context word with "like" when used as a preposition, conjunction, or interjection.

8.1.2. Cue assessment

Discriminant analyses were performed on the set of 1000 words using just the phonological cues, just the distributional cues, and the combined phonological and distributional cues. Analyses were weighted by frequency (so for the "like" example, the usage as a verb had a greater influence on the analysis results than did the other usages). Words were distinguished in terms of whether they were open or closed class words, and whether they were nouns or verbs. The random baseline was computed in the same way as for Experiments 1, 2, and 3, by randomly reassigning the grammatical category labels to words, but preserving the token frequency of each grammatical category.

8.2. Results

8.2.1. Open and closed class words

The discriminant analysis for phonological cues alone resulted in 84.3% correctly classified (53.9% random baseline), $\lambda = .457$. For distributional cues alone, 88.1% were correctly classified (53.5% random baseline), $\lambda = .359$. Finally, combined cues resulted in better classification than either cue type alone: 92.6% correct compared to 61.5% random baseline, $\lambda = .252$, all $p < .0001$. These results are similar to those of the disambiguated corpus in Experiment 3, in showing a very large contribution to correct classification from each cue type and a benefit of combined cues. For the individual cue types, classification increased in accuracy, whereas it dropped a little for the combined cues, though the improvement over chance classification remained at the same level. Comparing the correct classifications of the phonological cues and the distributional cues, the three-way hierarchical regression term was required to fit the data, indicating that, as with the previous analyses, the cue types were classifying distinct sets of words. Omitting the three-way term resulted in a highly significant mismatch between the model and the data, $\chi^2(1) = 331837.413$, $p < .001$, a replication of the effect of the disambiguated corpus in Experiment 3.

8.2.2. Nouns and verbs

70.0% of the words were correctly classified by the phonological cues (53.8% baseline), $\lambda = .748$. For the distributional cues, 83.7% were correctly classified (52.9% baseline), $\lambda = .554$. For the combined cues, 82.2% were correctly classified (57.1% baseline), $\lambda = .499$, all $p < .0001$. The results were very similar to the disambiguated corpus (Experiment 3) in the size of the improvement over chance, though distributional cues and combined cues were slightly lower in overall accuracy (compare Fig. 5). Omitting the three-way term from

the hierarchical loglinear regression for classifications made by the phonological cues alone and the distributional cues alone resulted in a significant difference between the model and the data, $\chi^2(1) = 147260.673$, $p < .001$, suggesting that the cue types operated on different sets of words in the corpus, again the same result as for the disambiguated corpus.

8.3. Discussion

The analyses based on a POS-tagged version of the corpus resulted in qualitatively similar results to the analyses in Experiment 3 where words were assigned their most frequent usage. For the open/closed class distinction, phonological cues alone were significantly better than chance, but not as effective as distributional cues alone, which in turn was less effective than combining both types of cue. For the noun/verb distinction phonological cues were once again found to be better than chance for classification, but not as effective as the distributional cues. The combined cues were found to be slightly less effective than the distributional cues alone, but with very similar levels of improvement over chance levels. Equally, the three-way hierarchical loglinear analyses indicated that the different types of cue were operating on different words within the lexicon, resulting in the three-way interactions, as in the original analyses.

Inspection of the words with multiple meanings showed that most ambiguous words had one usage that was much more frequent than its other usages. The pattern for the word "like", where the verb usage is most frequent, is reflected in most of the 1000 most frequent words in the corpus, and a consequence is that words that are ambiguous with respect to their class have only to be distinguished from their modal usage a small proportion of the time. For the word "like", the phonological cues assigned the word its most frequent usage—as an open class verb. However, the distributional cues were able to disambiguate the usage, such that the correct categorization for "like" when used as a preposition or a conjunction as well as when used as a verb was determined by the cues.

The POS-tagged corpus analyses indicated that the PDCH pattern of results were maintained when ambiguity of category usage was taken into account in the corpus of child-directed speech in English, indicating that assigning the most frequent category to each word is a valid simplification for analysis. However, the insight into cue use for ambiguous words indicates an additional advantage of multiple cue integration for determining multiple usages of words in the child's language.

9. Experiment 5: effects of morphology on classification

All the languages we have investigated have morpho-syntactic markers within the word. Hence, inflectional and derivational morphology could be the primary sources of the effects reported in the previous experiments, instead of their derivation from the phonological properties of word stems. MacWhinney and Bates (1989) suggested a "competition" hypothesis whereby languages with fewer constraints of word order tend to have more case-marking. A consequence of this would be that word-internal information would be observable when word-external information was weaker *across languages*, though this could not account for why there is an interaction between word-internal and word-external cues within a language for different grammatical categories. However, it remains to be determined to what extent the phonological properties of lexical categories are subsumed under morphology, and to what extent they are due to the more subtle properties of word

roots themselves, the properties noted by Kelly (1992) and Monaghan et al. (2005) for English. By focusing on roots, and stripping away morphological structure, we can assess whether the phonological cues to linguistic categories are mediated by morphological structure. It is, of course, possible that root forms, even though now morphologically inert, are to some degree reflections of prior morphological status. Indeed, a fundamental process of lexical change is the erosion of morphological affixes and their eventual collapse into the root word (e.g., Hopper & Traugott, 1993). Thus, past morphology might potentially have a small residual correlation with grammatical category. If this explanation is correct, we might expect that phonological correlations with grammatical categories should be present, but at a much reduced level. If, by contrast, morphology is not a major driving factor underlying the correlation, we should expect that highly significant correlations between phonology and grammatical category should remain, even when morphological structure has been stripped away.

9.1. Method

9.1.1. Corpus preparation and analysis

To test the effect of inflectional and derivational morphology on the cues for categorisation, we extracted all the monomorphemic words from the 1000 words in the English corpus from Experiment 1. We also included all the irregular verbs and noun plurals where morphology was not apparent from the surface form of the word (e.g., *knew, got, men*[6]). There were 822 words, of which 86 were classified as closed class and 598 were open class. There were 321 nouns and 152 verbs. We repeated the analyses precisely as before, using discriminant analyses to assess each cue type separately and combined, and determined a random baseline for the analyses.

9.2. Results and Discussion

For the open/closed class distinction, the phonological cues resulted in 67.8% correct classification, which was significantly better than the random baseline (63.0%), $\lambda = .602$. For the distributional cues, 90.0% were correctly classified, again better than chance (56.8% random baseline), $\lambda = .249$. For the combined cues, 95.2% were correctly classified (72.4% random baseline), $\lambda = .169$, all $p < .0001$. The three-way interaction between open/closed class category, phonological cue classification, and distributional cue classification was again significant, omitting this interaction resulted in a model significantly different from chance, linear regression $\chi^2 = 55{,}274.529$, $p < .0001$.

For the noun/verb distinction, the effects of the cues were again significantly better than the random baseline. For phonological cues alone, 68.0% were classified correctly (63.0% random baseline), $\lambda = .622$. For distributional cues alone, 94.9% were correctly classified (67.4% random), $\lambda = .235$. Finally, combined cues classified correctly 95.1% of nouns and verbs (76.4% random), $\lambda = .193$, all $p < .0001$. Omitting the three-way interaction from the hierarchical loglinear regression model resulted in a significant change from a good model fit, linear regression $\chi^2 = 8653.501$, $p < .0001$.

[6] Some of these forms are related to the vowel-change marker of past-tense in Proto-Indo-European (Chomsky & Halle, 1968), though no particular vowels were found to consistently mark a grammatical category in the current analyses.

The results of these monomorphemic analyses indicate that the interaction between different cue types is robust, and does not only depend on inflectional and derivational affixes in English. The effect of phonological properties on categorization, though slightly reduced compared to Experiment 3, nevertheless contributed significantly to classification. Another contribution to this reduction in the influence of phonological cues and increase in benefit of distributional cues was that monomorphemic words tended to be higher frequency on average than the whole corpus. Monaghan et al. (2005) demonstrated that higher frequency words followed this pattern of greater distinction between grammatical categories in terms of distributional properties than phonological cues, but that this effect reversed for lower frequency words. However, in short, the results of the monomorphemic analysis indicate that the role of different cue types in reflecting the grammatical categories of English cannot be explained in terms of the competition hypothesis of MacWhinney and Bates (1989).

10. General discussion

This paper presents accumulating evidence for the PDCH. The PDCH's central tenet is that every language will contain phonological cues that assist in determining the syntactic category of the word. Furthermore, the PDCH suggests that these phonological cues will operate in concert with word-external cues to determine a word's grammatical role. When external cues, such as distributional information, are weaker for a particular word, then phonological cues will typically be more reliable. The analyses of internal and external cues in English, Dutch, French, and Japanese have provided support for both these properties as potential language universals.

All of the four languages we have investigated incorporated a range of phonological cues that were significantly differently distributed across open and closed class words, and across nouns and verbs. This analysis of phonological properties provides some support to our use of such basic phonological properties of words, such as manner and place features as being a potential source for phonological coherence within a syntactic category, as indicated in Experiment 1. The discriminant analyses in Experiment 3 showed that these phonological properties could provide extremely accurate classification of open and closed class words (over 90% of open/closed class words correctly classified in Japanese, and over 90% of nouns/verbs correctly classified in Dutch). In all languages tested, the advantage of using phonological cues over a random baseline distribution were highly significant and demonstrated an effect size in the range 20–30% for Dutch, French and Japanese, and a smaller but still highly significant effect in English (5.6% for open/closed, and 10% for nouns/verbs). This effect meant that still over 50,000 words per million were correctly classified at a rate better than chance in English.

The distributional analyses we performed were also demonstrated to be language-general, and reflected significant differences between the occurrence of open/closed and noun/verb categories in each language. In all languages, the distributional information was extremely accurate in categorizing words, as shown in Experiment 3. The classifications were highly significant, and in the range 20–30% above the random baseline for each language. The achievement of over 90% correct classification based on this information is an indication of the value of contextual cues to syntactic category in every language, regardless of variations in word order.

Despite this high accuracy using distributional cues alone, the combined analyses, in all but one case, resulted in better performance than using one type of cue alone. Hence, the phonological and distributional cues categorized different regions of the lexicon. The complementarity of these cues was confirmed by the three-way loglinear analyses, indicating that the cue types were not overlapping to a significant degree. Figs. 4 and 6 indicated that phonological cues were more accurate at classifying verbs and closed class words than nouns and open class words, respectively. For the verbs and closed class words, distributional cues were less effective, and it is in the absence of this reliable information that the phonological cues are more evident. The results of this paper demonstrate that, using very language-general characterizations of phonological and distributional information, the role of the different cue types can be demonstrated as general across four very different languages, providing converging evidence for the PDCH.

The PDCH, at least in its strongest form, is a clearly falsifiable hypothesis, and in this study we have provided many opportunities for its falsification. We selected languages with very different properties—with variations in phonological properties (for example Japanese has no consonant clusters)—and variations in constraints on word-order. These variations provided a stringent test of the role of the types of information in categorization. Yet, remarkably similar patterns of effects were found in all four languages. Providing irrefutable evidence for the PDCH, just as with any other putative language universal, would require testing every extant natural language with a similar framework, which is clearly infeasible. To date, then, we have shown only that in selections from a range of languages—Germanic, Romance, and Japonic—the PDCH receives support. The strongest version of PDCH, and the one that we propose here, is that it corresponds to an exceptionless universal property of all languages; a weaker claim, of course, would be that it is a so-called statistical universal, that holds for most, but not all, languages (Hawkins, 1988). As is standard in linguistic methodology, it seems methodologically reasonably to propose the stronger claim, in the absence of any counterexample.

The PDCH suggests that acquisition of syntactic categories is facilitated by the correspondence between the phonological properties of words with the same sound. We suggest that any sensible mechanism for acquiring language will be sensitive to all available information, particularly if the information is potentially valuable for assisting in correct classification of up to 90% of words. The rational analysis approach to cognition (Anderson, 1994; Chater & Oaksford, 1999) contends that information that is useful for a task (and that is perceptually available) will be incorporated into the learning process. Consistent with this perspective, Fitneva, Christiansen, and Monaghan (submitted for publication) found that 7-year-olds used phonological cues when guessing whether a new word referred to a picture of an object or an action in a word learning task (see also Cassidy & Kelly, 1991, 2001). Furthermore, the phonological properties of syntactic categories appear to be invested in lexical access in adults, imparting an influence in single word naming and lexical decision (Monaghan, Chater, & Christiansen, 2003) as well as on-line sentence processing (Farmer, Christiansen, & Monaghan, 2006).

The beneficial use of phonological properties in language processing has been shown in numerous artificial and natural language learning studies (Cassidy & Kelly, 1991, 2001; Curtin, Mintz, & Christiansen, 2005; Mattys, White, & Melhorn, 2005; Saffran, Newport, & Aslin, 1996; Thiessen & Saffran, 2003), and some studies have indicated conditions under

which phonological coherence is required in order to learn statistical distributions of syllables (Onnis et al., 2005) or syntactic categories (Braine et al., 1990; Brooks et al., 1993). Given the added value of phonological coherence in each of the natural languages we have presented in this paper, it would be unsurprising that such information is used by the child in beginning the process of syntactic bootstrapping.

We have so far focused on the implications for language acquisition. But the complementary argument is also valid—that the lexicon, and language more generally, will tend to exhibit features that make it easier to acquire (e.g., Briscoe, 2002; Christiansen & Ellefson, 2002). This is because the lexicon is culturally transmitted across successive generations of language learners; and language properties which are easier to acquire will be transmitted more effectively. Thus, the lexicon would be expected to adapt, like the rest of language, through subsequent generations, to be easier to acquire. Such selectional pressures serve as a powerful mechanism which may establish and maintain phonological cues for categories.

Our results also have wider implications for the traditional conception of language and meaning. de Saussure's (1916) notion of the "arbitrariness of the sign" is often taken to mean that there is no systematic relationship between the phonological form of a word and how it is used. The phonological coherence across languages found in our analyses indicates that although the relationship between words and their individuated meaning may be largely arbitrary (with some exceptions in fragments of the lexicon—see Gasser, Sethuraman, & Hockema, 2005), there nonetheless is a systematic relationship between word forms and their syntactic category. Monaghan and Christiansen (2006) tested a set of computational simulations that mapped between pseudo-phonological representations of words and pseudo-meanings, and also between phonological representations and word categories. They found that the models learned more effectively when arbitrary relations existed for words between their phonological and meaning forms, but that systematic relations benefited learning the category to which the word belonged (see also Gasser, 2004). Hence, from a computational perspective, a language is most easily learned if it respects phonological coherence for syntactic categories but maintains as far as possible arbitrary meaning-form relations. The corpus analyses we have presented here demonstrate that the systematicity at the grammatical category level is very much in evidence in four diverse natural languages.

Thus far, we have focused on the commonalities across the different languages, but the analyses we have presented also indicate interesting differences between the languages. Perhaps most striking is that the phonological cues we have employed are effective to an equal degree in Dutch, French, and Japanese, but appear to be slightly weaker in English. This was found for both the open/closed class and the noun/verb distinction. This is consistent with Durieux and Gillis' (2001) analysis of Dutch using phonological cues inspired by English analyses, where they found better categorization for Dutch than English. It was perhaps surprising that Japanese demonstrated such a large effect of phonological coherence particularly for the open/closed class distinction, as the opportunities for expression of phonological similarity were more limited than in the other languages, given that morae begin with no more than one consonant. Japanese also demonstrated a surprisingly high degree of reliability in the distributional cues. French and English, where word order is less free, indicated the smallest effect of distributional cues for the open/closed class distinction, though

distributional cues in English and Dutch were the most effective for the noun/verb distinction.

The support for the PDCH from the current analyses invites some reappraisal of current computational models of language acquisition. For example, PARSER (Perruchet & Vintner, 1998) learns language structure by chunking items together that co-occur frequently, but does not take into account the phonological form of the chunks. Similarly, Cartwright and Brent (1997) and Redington et al. (1998) process words without regard to phonological similarity between words. Such models could be adapted by adding an additional computation that determines the similarity between the stimulus under consideration and other words or chunks that are already categorized by the process. Classifications based on simple recurrent networks already indicate the benefits of integrating cues from multiple sources. As mentioned above, Reali et al. (2003) demonstrated a neural network trained on information about phonological form and the position of the word in an utterance resulted in significantly better performance than just utterance position alone. It is this integrative nature of phonological information in categorization that we suggest is important in language acquisition, and that this information is not merely epiphenomenal or incorporated *post hoc* into processing.

The analyses of the PDCH reported here have established the presence of information in the child's language environment for supporting the development of grammatical categories. However, we have not specified a particular mechanism for how distributional and phonological information may be combined. Any self-organising system that learns to group together representations according to their similarity and that is responsive to similarities along several dimensions (e.g., distributional and phonological similarity) is likely to produce more accurate classification based on interacting sets of cues. One concrete starting point for how the PDCH may be instantiated in a language acquisition system is derived from the model of Redington et al. (1998). In their analyses, they used a clustering mechanism that grouped words together according to the similarities among their contexts. This system may be augmented by determining clustering according to both distributional and phonological similarity. Such a system may operate in tandem, or the initial clustering may be formed on the basis of one source of information alone and then refined by similarity from other sources of information. In the latter case, words which are distributionally dissimilar to other words in a phonologically-determined cluster would be reappraised or moved to another cluster where similarities among distributional representations are closer; or alternatively, phonological similarity could be incorporated directly into the measure of similarity used as a basic for building clusters of grammatical items. Approaches of this type would be consistent with the analyses of ambiguous words in Experiment 4, where categorization based on the phonological cues classified ambiguous words according to their most frequent usage, and then distributional information for lower-frequency uses was used to improve the classification of lower frequency usages for the word.

Another possibility is that distributional information may form the initial groupings, to be later filtered by phonological cues. In a study of the 5000 most frequent words in English child-directed speech, Monaghan et al. (2005) found that distributional cues were more reliable than phonological cues for high frequency words, whereas phonological cues were more reliable than distributional cues for lower frequency words. If

groups of words are formed based on the child's early experience, then these groupings are likely to be constructed for the words that the child has been exposed to more often. Hence, there is a richness of distributional information for these words, and less, though still evident, phonological information. Initial clustering based on distributional information rather than phonological information for these words would result in well-defined categories that respect the grammatical categories. As the categories develop to include rarer words, the distributional information is more likely to misclassify low frequency words. However, for these low frequency words the greater variation and reliability of the phonological information would have the effect of increasing the accuracy of these classifications. Whether phonological and distributional information operate serially or in parallel for defining categories, the analyses presented here indicate that accurate categorization can be achieved based on combining different cue types, and that these cue types interact in surprising ways.

The PDCH is consistent with the view that there are language universals; but we have argued that one source of such universals is that languages have been adapted to be easily learnable. To be learnable, the language must present with useable information about syntactic categories, and this information, when integrated across several sources, results in a system that is more easily learned. In contrast to nativist views of language universals, the universal property of phonological coherence is embedded within the communicative signal itself rather than being a property of the learner that is triggered by the stimulus (Pinker, 1999). Such a view does not preclude the possibility of an innate grammar, but the studies we have presented in this paper indicate the wealth of the stimulus in the language environment for the child, and the potential benefit of integration of information from multiple sources. We have only covered two potential sources of cues, and those only partially—there are, for instance, several cues shown to relate to category assignment in English phonology that were not included in the current analyses. Other sources of information to grammatical category, both within the speech signal, such as prosodic and allophonic information (e.g., Mattys et al., 2005), and objects and events in the environment co-occurring with certain utterances within the language (e.g., Yu & Smith, 2006), may provide additional reliable sources of information to grammatical category. The child's environment is a much richer source of information for language learning than we may have previously thought.

The evidence for multiple, universal phonological cues to grammatical category aligning with and complementing the distributional co-occurrence information in these languages is consistent with two theoretical views of language acquisition. It could be that all the requisite information for grammatical category learning in a particular language is present within the language signal, as in an empiricist view of language acquisition. Alternatively, it may be that the multiple cues work in tandem with innately specified semantic referents, as in the semantic bootstrapping account (Pinker, 1984). We suggest that, according to the principle of parsimony, a theory of language acquisition should first determine the potential of multiple cues within the language signal alone to constrain category learning, before hypothesizing innate structure that goes beyond the information contained in the language itself.

Appendix A

English:
he, we, one, don't, have, are, no, there, this, your, that's, on, in, oh, do, is, and, I, that, what, to, a, it, the, you.

Dutch:
ja, je, is, een, dat, de, het, wat, ik, niet, en, die, nou, maar, zo, dan, oh, ook, nee, in, he, wel, op, nog, moet.

French:
est, tu, c', qu', pas, la, a, ce, ça, le, oui, il, un, l', que, on, les, et, de, non, n', qui, je, des, dans.

Japanese:
doko, taro, sore, na, soo, ja, de, koko, ii, mo, nani, tte, da, n, ga, ka, a, ni, hai, wa, yo, kore, ne, un, no.

References

Anderson, J. R. (1994). *Rules of the mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Baker, M. C. (2001). *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

Bates, E., Devescovi, A., Pizzamiglio, L., D'Amico, S., & Hernandez, A. (1995). Gender and lexical access in Italian. *Perception and Psychophysics, 58*, 992–1004.

Bernstein Ratner, N., & Rooney, B. (2001). How accessible is the lexicon in Motherese? In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition;;* (Vol. 1, pp. 71–78). Amsterdam: John Benjamins.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.

Bowerman, M. (1973). Structural relationships in children's utterances: Syntactic or semantic? In T. Moore (Ed.), *Cognitive development and the acquisition of language*. Cambridge, MA: Harvard University Press.

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65–87). Hillsdale, NJ: Lawrence Erlbaum Associates.

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., et al. (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language, 29*, 591–610.

Briscoe, E. (Ed.). (2002). *Linguistic evolution through language acquisition*. Cambridge, UK: Cambridge University Press.

Brooks, P. B., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*, 79–95.

Campbell, R., & Besner, D. (1981). This and that—Constraints on the pronunciation of new written words. *Quarterly Journal of Experimental Psychology, 33*, 375–396.

Canavan, A., & Zipperlen, G. (1996). CALLHOME Japanese Speech. Linguistic Data Consortium, University of Pennsylvania.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*, 121–170.

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language, 30*, 348–369.

Cassidy, K. W., & Kelly, M. H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin and Review, 8*, 519–523.

Cassidy, K. W., Kelly, M. H., & Sharoni, L. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology: General, 128*, 362–381.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*, 57–65.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980). *Rules and representations*. Oxford: Blackwell.

Chomsky, N. (1981). Lectures on government and binding: The pisa lectures. Dordrecht: Foris Publications.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Ellefson, M. R. (2002). Linguistic adaptation without linguistic constraints: The role of sequential learning in language evolution. In A. Wray (Ed.), *Transitions to language* (pp. 335–358). Oxford, U.K.: Oxford University Press.

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

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Corbett, G. (1991). *Gender*. Cambridge, UK: Cambridge University Press.

Crain, S., & Lillo-Martin, D. (1999). *An introduction to linguistic theory and language acquisition*. Oxford: Blackwell.

Croft, W. (2003). *Typology and universals*. Cambridge, UK: Cambridge University Press.

Culicover, P. (1999). *Syntactic nuts*. Oxford: Oxford University Press.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96*, 233–262.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research, 22*, 109–131.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*, 133–142.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24*, 381–410.

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

Desrochers, A., Paivio, A., & Desrochers, S. (1989). L'effect de la fréquence d'usage des noms inanimés et de la valeur predictive de leur terminaison sur l'identification du genre grammatical. *Revue Canadienne de Psychologie, 43*, 62–73.

Dixon, R. M. W. (1977). Where have all the adjectives gone? *Studies in Language, 1*, 1–80.

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

Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, pp. 189–229). Amsterdam: John Benjamins.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences lexical processing. *Proceedings of the National Academy of Sciences, 103*, 12203–12208.

Fernald, A., & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. L. Morgan & K. Demuth (Eds.), *From signal to syntax* (pp. 365–388). Mahwah, NJ: Lawrence Erlbaum Associates.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the 14th annual meeting of the cognitive science society* (pp. 820–825). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. L. Morgan & K. Demuth (Eds.), *From signal to syntax* (pp. 343–363). Mahwah, NJ: Lawrence Erlbaum Associates.

Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (submitted for publication). From sound to syntax: Phonological constraints on children's lexical categorization of new words.

Fries, C. C. (1952). *The structure of English: An introduction to the construction of English sentences*. New York: Harcourt, Brace & Co.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39*, 218–245.

Gasser, M. (2004). The origins of arbitrariness in language. In *Proceedings of the cognitive science society conference* (pp. 434–439). Hillsdale, NJ: LEA.

Gasser, M., Sethuraman, N., & Hockema, S. (2005). Iconicity in expressives: An empirical investigation. In S. Rice & J. Newman (Eds.), *Experimental and empirical methods*. Stanford, CA: CSLI Publications.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development: Language, culture, and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gerken, L. (2001). Signal to syntax: Building a bridge. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, pp. 147–165). Amsterdam: John Benjamins.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249–268.

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

Hausser, R. (1989). *Principles of computational morphology. Technical report, laboratory for computational linguistics*. Pittsburgh, PA: Carnegie Mellon University.

Hawkins, J. A. (Ed.). (1988). *Explaining language universals*. Oxford: Blackwell.

Höhle, B., Weissenborn, J., Schmitz, M., & Ischebeck, A. (2001). Discovering word order regularities: The role of prosodic information for early parameter setting. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, pp. 249–265). Amsterdam: John Benjamins.

Hopper, P., & Traugott, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

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.

Kelly, M. H. (1996). The role of phonology in grammatical category assignment. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 249–262). Mahwah, NJ: Lawrence Erlbaum Associates.

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

MacWhinney, B., & Bates, E. (1989). *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. 2, pp. 127–214). New York: Gardner Press.

Marchand, H. (1969). *The Categories and types of present-day English word-formation* (2nd ed.). Munich, Federal Republic of Germany: C.H. Beck'sche Verlagsbuchhandlung.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*, 477–500.

McCawley, J. D. (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.

Mills, A. E. (1986). *The acquisition of gender: A study of English and German*. Berlin: Springer-Verlag.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition, 30*, 678–686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91–117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393–424.

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. 963–968). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Monaghan, P., & Christiansen, M. H., (2004). What distributional information is useful and useable in language acquisition? In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 963–968). Mahwah, NJ: Lawrence Erlbaum.

Monaghan, P., & Christiansen, M. H. (2006). Why form-meaning mappings are not entirely arbitrary in language. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1838–1843). Mahwah, NJ: Lawrence Erlbaum Associates.

Morgan, J. L., & Demuth, K. (1996). Signal to syntax: An overview. In J. Morgan & K. Demuth (Eds.), *From signal to syntax* (pp. 1–22). Mahwah, NJ: Lawrence Erlbaum Associates.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19*, 498–550.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20*, 67–85.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66*, 911–936.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of grammatical categories. In J. L. Morgan & K. Demuth (Eds.), *From signal to syntax* (pp. 263–283). Mahwah, NJ: Lawrence Erlbaum Associates.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique, 101*, 447–462.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Onnis, L. & Christiansen, M. H. (in press). Lexical categories at the edge of the word. *Cognitive Science*.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in speech processing. *Journal of Memory and Language, 53*, 225–237.

Perneger, T. V. (1998). What is wrong with Bonferroni adjustments? *British Medical Journal, 136*, 1236–1238.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning non-adjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General, 133*, 573–583.

Perruchet, P., & Vintner, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263.

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

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

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19*, 9–50.

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 Associates.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425–469.

Redington, M., Chater, N., Huang, C., Chang, L. -P., Finch, S., & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Chinese. *Proceedings of the Cognitive Science of Natural Language Processing*. Dublin.

Roach, P., & Hartman, J. (1997). *The English Pronouncing Dictionary* (15th Ed.). Cambridge: Cambridge University Press.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Sagae, K., MacWhinney, B., & Lavie, A. (2004). Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments and Computers, 36*, 113–126.

Sereno, J. A., & Jongman, A. (1990). Phonological and form-class relations in the lexicon. *Journal of Psycholinguistic Research, 19*, 387–404.

Sereno, J. A., & Jongman, A. (1995). Acoustic correlates of grammatical class. *Language and Speech, 38*, 57–76.

Shi, R. (1995). *Perceptual correlates of content words and function words in early language input*. Ph. D. Dissertation, Brown University, Providence, RI.

Shi, R., Morgan, J., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*, 169–201.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 27*, B11–B21.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86–132.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Statistical and stress cues in infant word segmentation. *Developmental Psychology, 39*, 706–716.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Boston, MA: Harvard University Press.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71–86.

Vance, T. (1987). *An introduction to Japanese phonology*. Albany, NY: SUNY Press.

Wolff, J. G. (1988). Learning syntax through optimisation and distributional analysis. In Y. Levy, I. M. Schlesinger, & M. D. S. Braine (Eds.), *Categories and processes in language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Yu, C., & Smith, L. B. (2006). Statistical cross-situational learning to build word-to-world mappings. In *Proceedings of the 28th annual meeting of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum Associates.