Available online at www.sciencedirect.com
SCIENCE DIRECT®
COGNITION
ELSEVIER
Cognition 96 (2005) 143–182
www.elsevier.com/locate/COGNIT

The differential role of phonological and distributional cues in grammatical categorisation

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

[a]*Department of Psychology, University of Warwick, Coventry CV4 7AL, UK*
[b]*Institute for Applied Cognitive Science, University of Warwick, Coventry CV4 7AL, UK*
[c]*Department of Psychology, Cornell University, Ithaca, NY 14853, USA*

Received 30 April 2003; revised 9 January 2004; accepted 13 September 2004

Abstract

Recognising the grammatical categories of words is a necessary skill for the acquisition of syntax and for on-line sentence processing. The syntactic and semantic context of the word contribute as cues for grammatical category assignment, but phonological cues, too, have been implicated as important sources of information. The value of phonological and distributional cues has not, with very few exceptions, been empirically assessed. This paper presents a series of analyses of phonological cues and distributional cues and their potential for distinguishing grammatical categories of words in corpus analyses. The corpus analyses indicated that phonological cues were more reliable for less frequent words, whereas distributional information was most valuable for high frequency words. We tested this prediction in an artificial language learning experiment, where the distributional and phonological cues of categories of nonsense words were varied. The results corroborated the corpus analyses. For high-frequency nonwords, distributional information was more useful, whereas for low-frequency words there was more reliance on phonological cues. The results indicate that phonological and distributional cues contribute differentially towards grammatical categorisation.

Keywords: Language acquisition; Syntactic categorization; Phonological cues; Distributional information

* Corresponding author. Address: Department of Psychology, University of York, York YO10 5DD, UK. Tel.: +44 1904 432885; fax: +44 1904 433181.
E-mail address: pjm21@york.ac.uk (P. Monaghan).

0022-2860/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2004.09.001

1. Introduction

A necessary prerequisite to producing sentences is that the language learner derives a knowledge of the different grammatical categories and the relations between them. Knowing the category of a word is also a precursor to understanding referents in other's speech. Given the importance of this knowledge in language acquisition it is not surprising that so much debate has centred on this issue, particularly over how grammatical category information is attained. At one level, discussions have concerned whether the categories themselves are innate (Pinker, 1984), or can be learned (though it is, of course, agreed that assignment of lexical items to categories is learned). Assuming that grammatical categories can be learned, another level of debate concerns the sources available to the child in order to learn such categories. Explanations have been offered that invoke the importance of semantic (Bowerman, 1973; Macnamara, 1972), phonological (Kelly, 1992), and distributional (Harris, 1951) cues in the learning process. These have been reviewed in detail elsewhere and so we do not consider them at length here (Christiansen, Allen, & Seidenberg, 1998; Christiansen & Dale, 2001; Mintz, Newport, & Bever, 2002; Redington & Chater, 1998). Several studies have explored the potential value of using one type of cue, either phonological or distributional, yet the benefits of integrating information between the different types has not been assessed empirically. This paper provides a test of how information is integrated across these different modalities of cues, employing corpus analyses of child-directed speech and an artificial language learning experiment.

2. Cues for grammatical categorisation

There are numerous studies that have assessed phonological and distributional information in determining the grammatical category of words. We review these in turn.

2.1. Phonological cues in grammatical categorisation

Several studies have indicated that phonological cues are either *useful* or *used* for grammatical categorisation. Studies investigating the potential usefulness of phonological cues have typically consulted corpora to indicate the different distributions of cues for separating grammatical categories. Such cues can be classified in terms of whether they refer to phonological properties at each of three levels: the word level, the syllable level, and the phoneme level. We located 16 such cues in the literature. As these cues form the basis of our phonological corpus analyses, we report our encoding scheme for each cue in parentheses.

At the word level:

1. *Length in phonemes*: open class words are generally longer than closed class words (Morgan, Shi, & Allopenna, 1996), and nouns are generally longer than verbs (Kelly, 1992). (In our scoring scheme, we counted the number of phonemes in each word.)
2. *Length in syllables*: closed class words have a minimal number of syllables, and may even be subsyllabic (e.g. *he is* may be contracted to *he's*), which is consonant with

Morgan et al's (1996) theory of perceptual minimality in closed class words. Also, nouns have more syllables than verbs (Kelly, 1992). (We counted the number of syllables in each word.)

3. *Presence of stress*: words with no stress are more likely to be closed class than open class (Gleitman & Wanner, 1982). (Words that were not stressed scored 1 and all words with stress scored 0.)
4. *Position of stress*: words with iambic stress (stress on second syllable) are more likely to be verbs, whereas words with trochaic stress (first syllable stress) are more likely to be nouns. From an assessment of 3000 disyllabic nouns and 1000 disyllabic verbs, 90% of the words with first syllable stress were nouns, and 85% of the words with second syllable stress were verbs (Kelly & Bock, 1988). (In our scoring scheme, words scored 0 if they had no stress, 1 if primary stress was on the first syllable, 2 if primary stress was on the second syllable, and 3 if primary stress occurred later in the word).

At the syllable level:

5. *Onset complexity*: open class words are more likely to have consonant clusters in the onset than closed class words (Shi, Morgan, & Allopenna, 1998). (We counted the complexity of the onset in terms of number of consonants it contained from 0 (for words beginning with a vowel), to 3 (e.g. /stɹ-)
6. *Word complexity*: open class words are more likely to have consonant clusters in the onset and codas of all syllables than closed class words (Morgan et al., 1996). (We measured the proportion of the phonemes that were consonants in each word.)
7. *Proportion of reduced vowels*: closed class words are more likely than open class words to appear in reduced form (Cutler, 1993). (For each word, we counted the proportion of the syllables in each word that were pronounced either as /ə/or were syllabic consonants (e.g. the /l/ in bottle). Words with no vowels took a value of 1.)
8. *Reduced first syllable*: closed class words are more likely than open class words to have a reduced vowel in the first syllable (Cutler, 1993; Cutler & Carter, 1987). (We measured whether or not the first syllable in each word was reduced (/ə or a syllabic consonant. If there were no syllables then it scored as if the first syllable was reduced.)
9. *-ed inflection*: adjectives are more likely than other words to end with syllabified "-ed": *ragged* is pronounced/ɹ-ægɪd/ as an adjective, but/ɹ-æɪgd/as a verb, and the end of *learned* can be pronounced/əd/as an adjective, whereas this pronunciation is not permitted when the word is used as a verb (Marchand, 1969, cited in Kelly, 1992). (The word scored 1 if the last syllable was composed of a consonant or consonant cluster followed by/ə d/or /ɪd/, and scored 0 otherwise.)

At the phoneme level:

10. *Coronals*: Morgan et al. (1996) showed that closed class words are more likely to contain coronal consonants (/d/, /t/, /ə/, /ð/, /s/, /z/, /d/ʤ/, /ʧ/ʃ/, /n/, /l/, or /ɹ-/ than open class words. (We counted the proportion of consonants in each word that were coronals. Words with no consonants took a value of 0.)
11. *Initial /ð/:* closed class words are more likely to begin with the voiced alveolar plosive /ð/ than open class words (Campbell & Besner, 1981). (Words beginning /ð/ scored 1, and all other words scored 0.)

12. *Final voicing*: if a word finishes in a final consonant, this is more likely to be voiced if the word is a noun rather than a verb (Kelly, 1992). (The word scored 1 if it ended with a voiced consonant, scored 2 if it ended with an unvoiced consonant, and scored 0 if it ended with a vowel.)
13. *Nasals*: nouns are more likely than verbs to contain nasals (Kelly, 1992). (We scored the proportion of consonants in the word that were nasals. Words with no consonants scored 0.)
14. *Stressed vowel position*: vowels in stressed syllables tend to be back vowels in nouns and front vowels in verbs. This was shown to be the case for high frequency words (Soreno & Jongman, 1990). (We assessed the position of the vowel in the syllable with primary stress. If the vowel was a front vowel, it scored 0, a central vowel scored 1, and a back vowel scored 2. Words with no primary stress were coded as central vowels. For words containing a glide or a diphthong, we averaged the position of the vowels, so the diphthong /iə/ scored 0.5 as /i/ is a front vowel, and /ə/ is a central vowel.)
15. *Vowel position*: vowels in nouns tend to be back vowels, and in verbs tend to be front vowels. (We assessed the mean position of vowels in the word, on a scale of 0 for all front vowels, to 2 for all back vowels. Words with no vowel were scored with a mean central vowel position.)
16. *Vowel height*: vowels in nouns tend to be low, and vowels in verbs tend to be high. (We encoded the mean height of vowels in the word, on a scale of 0 for all close vowels to 3 for all open vowels.)

Two recent studies have explored the benefits of combining different phonological cues for grammatical categorisation (Shi, Morgan & Allopenna, 1998; see also Morgan et al., 1996) assessed a number of phonological cues to distinguish open and closed class words in small corpora of child-directed speech (<100 words for tests of Mandarin speech, and 200 words for their analysis of Turkish). They assessed words for three cues at the word level of analysis: type frequency, utterance position (initial, medial or final), and number of syllables; three cues at the syllable level of analysis: presence of diphthongs, presence of syllable coda, and syllable duration. They also assessed two pitch cues: relative amplitude, and rate of pitch change, and language-specific cues (for Mandarin, syllable reduplication, to add delimitation to verbs or add vividness to adjectives or adverbs, and presence of marked tone. For Turkish, vowel harmony). They then trained a Kohonen network to map the cues to provide a two-dimensional representation of the words.

The Kohonen network was trained on 60% of the words from the Mandarin corpus and then tested on the remaining 40% of the words. Words were classified correctly if they produced activity in units in an area of the network that was also activated for other words of the same class. Twenty percent of the words were not classified, but of those that were classified approximately 90% of the words were correctly assigned to open or closed class category. For the Turkish speech, results were comparable, with all but one cue significantly differently distributed between open and closed class words (pitch change was not significant). A Kohonen network, trained and tested in a similar way as for the Mandarin analysis, resulted in no classification for approximately 20% of words, and correct classification for 80–85% of the words that were classified.

Shi et al's (1998) study provides a small-scale but impressive display of the value of combining multiple cues for grammatical categorisation of open and closed class words. The second study combining multiple cues aimed to distinguish nouns and verbs taken from a large dictionary with phonological forms for each word. Durieux and Gillis (2001) assessed the usefulness of a set of cues proposed by Kelly (1996) for distinguishing nouns from verbs. A set of nouns and verbs was randomly sampled from the CELEX English database in such a way that both nouns and verbs of all frequencies were selected. They employed cues measuring position of stress, vowel height, presence of nasals in onset, presence of nasals in coda, and number of phonemes per syllable. The model used to assess the cues was based on instance-based learning, which stored examples in memory and compared new items to those stored items. The classification of the new word was taken to be the same as that of the closest stored item.

For nouns and verbs, the combined cues resulted in correct classification of 77.59% of nouns and 61.37% of verbs. When all open class categories were considered, the model correctly classified 76.30% of nouns, 71.48% of verbs, 50.58% of adjectives, and 80.09% of adverbs. The same set of cues was found to distinguish nouns and verbs in Dutch, to greater accuracy than in English (81.77% of nouns and 67.97% of verbs). Durieux and Gillis (2001) also performed a "phonological encoding" analysis where cues were not defined *a priori*, but rather words were represented in terms of features for each onset, nucleus and coda for each syllable. In this analysis, 79.24% of nouns and 75.70% of verbs were correctly classified, and when stress was also added performance increased to 84.18% of nouns and 79.41% of verbs. Words of different frequencies were analysed for correct classification, and lower frequency words were found to be classified more accurately, indicating that phonological and distributional cues may play a different role for words of high and low frequency. They attributed poorer performance for higher frequency items as being due to the greater ambiguity of high frequency words with respect to grammatical category. High frequency words are more likely to occur as both nouns and verbs, for example. In addition, they indicated that there were proportionally more nouns in the lower frequency sets, and nouns were classified with greatest accuracy in the other analyses. Correct classifications in the instance based learning model are more likely if there is a predominance of one category (i.e. the random baseline for performance will increase alongside increasing accuracy for the actual data).

The above studies have indicated the *usefulness* of phonological cues for grammatical categorisation. In addition, there are a set of studies that have probed the *use* made of phonological cues in categorisation. Cassidy and Kelly (1991) required adult participants to place a nonword in a sentence context. If the nonword was of one syllable in length then it was more likely to be used in a verb context, whereas if it was three syllables in length then it was more likely to be used as a noun. Cassidy and Kelly (1991) claimed that the results indicated that participants were sensitive to the phonological cue of syllable length that distinguishes nouns from verbs. Studies on children who heard a nonword and had to point either to an action or an object in a picture (Cassidy & Kelly, 1991), or relate the nonword to an action or an object in a short videoed scene (Cassidy & Kelly, 2001) produced similar results. When the nonword was one syllable in length the children tended to point to the action, and for the three syllable nonwords, they tended to point to the object. Though individual cues seem to be highly unreliable when considered alone,

the results of these studies suggest that cues can be employed individually for determining the category of novel words.

2.2. Distributional cues in grammatical categorisation

A number of studies have addressed the issue of distributional information in grammatical categorisation (Bloomfield, 1933; Finch & Chater, 1992; Fries, 1952; Harris, 1954; Kiss, 1973; Maratsos & Chalkley, 1980; Schütze, 1993; Wolff, 1988). Redington, Chater, and Finch (1998) provided a detailed illustration of the potential value of distributional information in providing evidence of grammatical category. They assessed the local contexts of the most frequent 1000 words in the CHILDES corpus of transcribed child-directed speech. For each word, its co-occurrence with the 150 most frequent words was counted at positions one before, one after, two before, and two after. The four resulting context vectors were combined to produce a 600-dimensional vector. The co-occurrence vectors for words were compared, and clustered together according to similarity using hierarchical cluster analysis. Words were labelled with their grammatical category, and the objective categories were compared to those produced by the cluster analysis. The results of the classification were good, with best performance resulting from a cut-off of similarity at level 0.8. At this cut-off, 72% of words of the same category were accurately clustered, with 47% completeness of classification (compared to a random baseline of 27 and 17%, respectively). When only nouns and verbs were considered, performance was even better. Nouns were clustered with accuracy 90% and completeness 53% (baseline 43 and 14%) and verbs were clustered with accuracy 72% and completeness 24% (baseline 25 and 14%).

Redington et al. (1998) analyses were particularly striking as they were unsupervised: information about category was not provided prior to construction of the clusters, category information was only used to assess the results of the clustering based on the distributional co-occurrence information. The point at which the hierarchical clusters have to be cut to produce the categories has to be decided upon, and this was performed in a supervised manner, in that the authors selected the cut-off that produced the best match to the objective grammatical categories. Yet, such results provide a benchmark for the extent to which information about grammatical category may be constructed without prior knowledge of the categories.

Mintz, Newport, and Bever (2002) performed a similar analysis to that of Redington et al. (1998), computing co-occurrence vectors and clustering words in terms of the similarity of their vector representations, except using small corpora of speech directed to very young children. They found that such input produced clusters of words of the same categories with an accuracy greater than chance.

An alternative method for assessing the value of distributional information in categorisation involves computing particular frames in which words from particular categories occur. Cartwright and Brent's (1997) model searched for pairs of sentences that differed minimally, i.e. in terms of differing over a single word, resulted in a grouping of the differing words into a category, and the abstraction of a template in which they occurred. Thus, the two sentences *I saw the cat* and *I saw the dog* would result in defining a template *I saw the N*, where *N* is composed of *cat* and *dog*. Their model was trained on the Bernstein–Ratner child-directed speech corpus (a subcomponent of the CHILDES corpus), and it performed at a level of 68.1% (22.6%) accuracy and 22.0% (22.6%) completeness, in terms

of correctly grouping grammatical categories together (random baseline values in parentheses). Increasing the size of the corpus did not improve completeness and resulted in a gradual decline in accuracy, so it seems unlikely that applying this framework to a significantly larger corpus would result in better performance.

Fries (1952) listed a set of frames in which words of different categories could (only) occur. For example, any word that could fill the gap in *(The) — is/was/are/were good* has to be a noun, where *the* in parentheses indicates that this is optional, and the slash indicates one word from the set of options. Fries identified a set of 19 such templates into which words of different categories fitted. Similarly, Maratsos and Chalkley (1980) considered the possibility that children use the local context of a word to determine grammatical category. They describe several frameworks within which a noun may occur but not a verb, and vice versa. For example, a verb may occur with the inflection *-ed*, whereas a noun may not. Such approaches indicate that the very local context of a word provides a great deal of information about its grammatical category.

Mintz (2003) provided an empirical test of the potential information available for classifying words into different categories when using frames similar to those employed by Maratsos and Chalkley (1980). He assessed a small corpus of child-directed speech for the occurrence of words in the 45 most frequent three-word frames (such as *The — is*), and found that classifications according to grammatical category were achieved with 93% accuracy and 8% completeness,[1] which were significantly higher than random baselines (47% accuracy and 4% completeness).

Trigram distributional information has therefore been indicated to be useful for categorisation when the target word is in central position. However, corpus studies of high-order *n*-gram statistics (such as in Cartwright & Brent, 1997; Mintz, 2003) demonstrate high accuracy but low completeness. If a word occurs in a complex frame then it is very likely to be of a particular category but the greater complexity of the frame entails that fewer instances of words will occur in that frame. Hence, lower-order distributional information may be useful in achieving a greater degree of completeness, but perhaps at the expense of accuracy (Monaghan & Christiansen, 2004). Bigram analyses, for instance, will categorise many more words (more nouns follow *the* than occur in the central position of *the — is*) but are more vulnerable to speech errors or false starts in speech.

Gómez (2002) found that longer-distance dependencies (such as trigrams) were only learned in artificial language experiments when the bigrams in the stimuli were uninformative, either because the bigram transitional probabilities were very high or very low. Her language consisted of sentences of three words in length, with the third word always predictable from the first word. When the intervening word was from a small set (so bigram frequency was high) the trigrams were not learned. When the variability of the middle word was high, so bigram frequency was low and uninformative, then the trigram structure could be learned. Onnis, Christiansen, Chater, and Gómez (2003) tested the special case where there was no variability for the middle word and again bigram statistics were uninformative, and found that trigrams could be learned. In segmentation studies,

[1] For analysis of word types under standard coding. Scores for token analyses, and for expanded encoding were broadly similar.

transitional frequencies at the bigram level were learned by both infants and adults in continuous streams of syllables (Aslin, Saffran, & Newport, 1996; Saffran, Aslin, & Newport, 1996). Mintz's (2002, 2003) studies indicate that categorisation can take place on the basis of trigram information, but his analyses do not preclude the contribution of learning at the level of bigrams.

Smith (1966, 1969) tested the extent to which bigram sequences could be learned. Participants were exposed to sentences of the form MN or PQ, where M was comprised of four words, as was N, P, and Q. Participants were then asked to recall sentences that they had heard. Participants produced pairs that respected the ordering, such as MQ and PN, in addition to reproducing sentences that they had heard. They did not produce pairs that violated orderings, such as NP or NQ pairs. However, categorisation based on this structure was not assessed, such as whether MN pairs were judged to be part of the language over MQ pairs, and so these studies do not provide evidence for or against learning of bigrams for categorisation. Foss and Jenkins (1966) provide evidence that categorisation can be learned from a similar language when the relative size of the sets M and P are distinct from those of the sets N and Q. They taught people to associate a set of words with one of two markers, and then tested their transfer of these groupings to associations with new markers. When the set size was 20 or 10, transfer performance was good, but no better than chance when the set size was just 6. Hence, categorisation appears to be better when there is large variability in the categorised set, compared to the set size of the context-word cues used for classification. Such a structure reflects the high frequency of closed class words such as "the" or "to" against the large variability of the words that can follow such items (nouns and verbs).

Valian and Coulson (1988) also provided an empirical test of the extent to which bigram distributional information could contribute towards categorisation. We report this test in detail as it provided the basis for the design of the artificial language learning experiment in this paper. The artificial language they constructed consisted of two categories *A* and *B*, where each category consisted of six words. Words within a category were always preceded by the same high-frequency marker word, so *A* words were always preceded by the word *a*, and *B* words were always preceded by the word *b*. Sentences were of the form *aAbB* or *bBaA*. Participants were trained on 24 such sentences, and then tested on 12 sentences that conformed to the language structure and 12 sentences that did not. Of the 12 incorrect sentences, three violated the ordering of the marker word and the category word (e.g. *aABb*); three violated the alternating marker-word/category-word structure (e.g. *aABB*); three violated the pairing of marker-words with the correct category word in one of the pairs (e.g. *aAaB*); and three violated the pairing of marker word and category word in both pairs (e.g. *bAaB*). The training and testing was repeated four times. Performance was compared to learning in a language where there were four marker words, two assigned to each category, and three words in each category. Valian and Coulson (1988) found that participants learned more quickly and to greater accuracy in the high-frequency condition, and this was due to differences in accuracy for learning violations of the third and fourth types (when marker-word and category-word were wrongly paired). The high-frequency words in the artificial language were interpreted as acting as anchor points around which the structure of the language could be determined. The frequency of the marker-words determined ease of learning of the language.

Distributional cues prove extremely useful for determining grammatical category. Greater specificity of context results in a greater degree of accuracy in categorisation, but with lower completeness than may be achieved by taking more general, lower-level contextual information into account. This is intuitively suggested by cutting the clusters resulting from Redington et al.'s (1998) study at different levels. Cutting at a low-level resulted in high accuracy but low completeness as there are many separate clusters. Cutting at a high-level resulted in low accuracy but high completeness as there are just a few clusters. Clustering at higher levels exploits increasingly general information in the distributional structure. Studies on artificial language learning suggest that structure at different levels of generality-both trigrams and bigrams-can be learned by participants.

3. Combining distributional and phonological cues

Shi et al.'s (1998) analyses may be interpreted as combining phonological, acoustic and distributional cues, in that frequency and utterance position could be considered to be distributional cues. The differences in distributions for each cue were significant in their study, but it remains unclear how much information each source contributed towards correct classification, and what benefits may accrue from combining information between sources. A number of issues remain unresolved by these previous studies on cue use in language acquisition.

First, previous studies of phonological cues in categorisation have either focused on very small corpora, or have not been informed by child-directed speech (Durieux & Gillis, 2001; Kelly, 1992; Shi et al., 1998). In Experiment 1 we provide a detailed analysis of the validity of phonological cues in grammatical categorization on a large corpus of child-directed speech. We employ all 16 phonological cues that we have identified in the literature.

Second, there have been no previous large-scale empirical tests of bigram co-occurrence statistics and their usefulness for categorization, in contrast to previous studies of distributional information that invoked longer-distance dependencies between words (e.g. Mintz, 2003; Mintz, Newport, & Bever, 2002; Redington et al., 1998). Experiment 2 tests the extent to which bigram information provides successful cues for grammatical categorization.

Third, little is known about how cues are integrated, particularly across different modalities (Christiansen & Dale, 2001). One possibility is that different cues will be useful for different situations. In particular, we hypothesise that distributional information will be more useful for categorising higher frequency words, whereas phonological information will provide more valid data for lower frequency words. This is because high frequency words are more likely to have reliable contextual information, but undergo compression in terms of their phonological form.

The prediction that distributional information will be most useful for higher frequency words stems from the claim that contextual information for a word becomes more reliable as more instances of a word are heard. If a word is heard only once, then it is possible that it may have occurred in error—child-directed speech, similar to adult-adult speech, is replete with false starts, single-word utterances, and ungrammatical

constructions (Lickley & Bard, 1998). A single token of a word, then, may provide misleading evidence regarding the use of that word in general. Additional occurrences of the word enable the hearer to increase the confidence of the reliability of the word's context. For example, hearing a word preceded by *the* once may give the listener a hint that the word is a noun. However, *the* is a highly frequent word and occurs in uninformative contexts many times in speech (*the* precedes *the* 393 times in the adult child-directed speech from the CHILDES corpus (MacWhinney, 2000), for instance, but *the* is not a noun). It would therefore be an accident-prone policy to categorise the word on the basis of the context of its first use. Yet, if the listener hears the same word several times and each time it is preceded by *the*, then the listener will begin to encode that is not a mere accident that the target word follows *the*. It follows, then, that the more instances of a word that are heard the greater is the certainty of the accumulated contextual information for that word as indicating the word's usage. We return to this point below, in the discussion of the distributional cues that we employ.

In contrast, we suggest that phonological cues will provide *less* information about higher frequency words than lower frequency words. This is because higher frequency words tend to be subjected to contractions and assimilations in the speech signal. High frequency usage results in a reduction of the physical signal of the word (Cutler, 1993), and Zipf's Law reflects this fact: high frequency words tend to be shorter in length (Zipf, 1935). The phonological forms of words for high frequency items, then, will tend to converge. If phonological cues correspond to different grammatical categories, the value of such cues will be less emphatic for these higher frequency items, because other forces have been brought to bear on the words, constraining them to be closer in terms of their phonological representation. Lower frequency items, on the other hand, are not prone to these forces of compression to the same degree. Differences in terms of phonology are more likely to be greater for these lower frequency items, and this would be a serendipitous feature given that distributional cues are not applicable for lower frequency items. In this respect, our hypothesis about different application of cues for low- compared to high-frequency words differs from that of Durieux and Gillis (2001). Phonological information is poorer for high-frequency words because of communicative pressures rather than greater category ambiguity or presence of more words of a particular category. We test this hypothesis in Experiments 1 and 2 for different frequency groupings, and also in Experiment 3 where we combine phonological and distributional cues in our analyses.

Each source of information—phonological or distributional—determines essential differences between words in terms of the different grammatical categories. Though other types of cue, e.g. semantic, are also undoubtedly useful, we concentrate on the extent to which categorisation can be successfully achieved without yet incorporating those data sources. The first three experiments are based on corpus analyses, which pursue a rational analysis approach towards language learning, in that they indicate the potential information available in the environment for grammatical categorisation. A computational system operating optimally will pick up on such signals. It is possible that the processes of language learning may, for some reason, be suboptimal, and so we also test the availability of such cues in a learning experiment. We show that the experimental results support the outcomes of the corpus analyses. The first experiment investigates

the potential role of phonological cues in distinguishing different grammatical categories, the second assesses the role of distributional cues, and the third combines both the phonological and the distributional cues. The fourth experiment tests the predictions raised by the corpus analyses in an artificial language learning study.

4. Experiment 1: testing phonological cues in grammatical categorization

4.1. Method

Corpus preparation. The corpus was derived from the CHILDES database of child-directed speech. We extracted all the speech by adults from all the English corpora in the database, resulting in 5,436,855 words. We replaced pauses and stops with boundary markers, producing 1,369,574 utterances in the corpus. The average length of an utterance was 3.97, which is in accordance with an assessment of the Bernstein Ratner fragment of the CHILDES corpus (Bernstein Ratner & Rooney, 2001). The CHILDES database is formed from recordings of the speech environment of children from various ages. We used all adult speech in the whole corpus for our analyses, in contrast to other studies that employed fragments of the corpus (e.g. Mintz et al., 2002), as the large scale was beneficial for determining the idealised input to the child-smaller corpora provide an impoverished reflection of the input for even young children. Smaller fragments, for instance will not provide an accurate reflection of the Zipf-like distribution of the language and this particularly affects the representativeness of the occurrence of lower frequency words.

The CHILDES database provides (with the exception of only a fragment of the database) only orthographic transcriptions of words,[2] so we derived phonological and syntactic category for each word from the CELEX database (Baayen, Pipenbrock, & Gulikers, 1995). Many words have alternative pronunciations, and so we took the most frequent pronunciation for each orthographic form. Many orthographic forms can also be used in more than one grammatical class—record, for example, can be used as a noun or as a verb, though with different pronunciations. Again, we took the most frequent syntactic category for each orthographic form, and so ignored the potential information in different pronunciations of such homographs. Selecting the most frequent phonological and grammatical form of each word contributes noise to the analysis and provides the weakest test of the contribution of these cues towards categorisation. Hand-coding words would be impractical for such a large corpus which is necessary for reliable estimates of the relationship between cues and the whole lexicon. Absence of a contribution for individual cues, for example stress position distinguishing nouns and verbs, could be due to the noise inherent in this encoding.

In the analyses below, we considered the most frequent 5000 words in the CHILDES database. Of these 5000 words, several did not have corresponding phonological forms in

[2] Brian MacWhinney has recently provided a parsed version of the entire English CHILDES database (http://childes.psy.cmu.edu/data/eng-uk-mor).

the CELEX database. We hand-coded the 72 words that occurred in the 1000 most frequent words from the CHILDES database but did not occur in the CELEX corpus. These were largely alternative spellings, such as *wanna* for *want to*, proper nouns, or interjections, such as *hmm*, or *arrgh*. We omitted from our analyses all of the other less frequent words that did not have phonological forms.

Cue derivation. The phonological form for each word was assessed for the 16 phonological cues, mentioned previously, which distinguish words at the whole word level, the syllable level, or at the phoneme level. The cues are summarized in Table 1, together with examples of the values taken by particular words in the analyses.

Statistical tests. We focus on two distinctions that have been considered in the literature: open class and closed class words, and, within the open class category, nouns and verbs. These distinctions were chosen because they represent large sections of the lexicon, and indicate the extent to which cues can be used to distinguish words with very different usages. Significant differences between the means of two classes in terms of an individual cue does not mean that the cue can be used reliably as a predictor. Morgan, Shi, and Allopenna (1998) point out that the distributions of two classes with different means may have substantial overlap, meaning that accurate classification using that cue is substantially close to the baseline of an entirely random classification. We therefore distinguish tests of *significance*, reflecting the significance of the distributions of the cues across categories, from tests of *diagnosticity*, which reflect the extent to which tests can accurately classify items into categories. Thus, we report the individual measures separately, and then report the combined contribution of each measure towards predicting class membership.

Table 1
Summary of cues used in the analyses

Cue number	Cue name	Examples
1	Phoneme length	Penguin (6), veer (3)
2	Syllable length	Stratification (5), there (1)
3	Presence of stress	Con'sume (1), a (0)
4	Stress position	Income (1), re'joinder (2)
5	Onset complexity	Street (3), in (0)
6	Syllabic complexity	Stress (0.8), calling (0.6)
7	Reduced syllables	Determine (0), the (1)
8	Reduced 1st vowel	Revoke (0), to (1)
9	-ed inflection	Faced (0), detested (1)
10	Coronal	Obscure (0.33), on (1)
11	Initial/ð/	Think (0), though (1)
12	Final voicing	Bead (1), bent (2)
13	Nasal	Man (1), find (0.33)
14	Stressed vowel position	Loading (1.5), moth (2)
15	Vowel position	History (1.33), offer (2.5)
16	Vowel height	Coffee (1.75), page (0.75)

Examples of words with high scores and low scores for each cue are indicated in the final column (numbers in parentheses indicate scores for the given words).

4.2. Results: distinguishing open- and closed-class words

We selected words classified as nouns, verbs, adjectives or adverbs in the CELEX database as open class words. Prepositions, conjunctions, articles and pronouns were classified as closed class words. We omitted numerals, interjections and contractions (e.g. *don't*, *would've*) from our analyses.

Tests of significance. We performed Mann–Whitney U-tests on the difference between the means of the 4569 open class words and the 146 closed class words in the most frequent 5000 words in the CHILDES corpus. The results of the tests are shown in Table 2. For word-level cues, open class words were more likely to have stress and, relatedly, more likely to have stress at a later syllable in the word. Open class words also had more phonemes than closed class words. For syllable-level cues, open class words had more complex onsets and syllables, closed class words had more reduced vowels and more reduced first vowels. At the phoneme level, closed class words began with //ð// more than open class words.

The tests of differences in means confirmed that, in large-scale analyses, several of the cues posited as useful for distinguishing open from closed class words by Morgan et al. (1996) were differently distributed across open and closed class words. However, taken alone, each cue is highly unreliable, even though classifications based on an individual cue may be highly significant, i.e. the diagnostic tests may not succeed even when significance tests are positive. We illustrate this below for nouns and verbs (the examples for open and closed class words are less immediately apparent due to the difference in size of the groupings). It is possible that different cues distinguish different groups of open and closed class words, and that, therefore, the classifications based on combined cues may increase accuracy.

Table 2
Comparisons between open and closed class words in the most frequent 5000 words for the 16 phonological cues

Phonological cue	Open class	Closed class	Z
Phoneme length	4.77	3.95	−6.057***
Syllable length	1.70	1.57	−2.191
Presence of stress	1.00	0.47	−49.489***
Stress position	1.10	0.75	−11.323***
Onset complexity	1.13	0.71	−8.896***
Syllabic complexity	0.66	0.59	−7.401***
Reduced syllables	0.12	0.10	−1.218
Reduced first vowel	0.03	0.10	−4.481**
-ed Inflection	0.01	0.00	−1.346
Coronal	0.61	0.65	−1.815
Initial/ð/	0.00	0.09	−18.713***
Final voicing	1.16	1.02	−2.570
Nasal	0.17	0.21	−0.129
Stressed vowel position	1.78	1.89	−2.370
Vowel position	0.73	0.77	−0.496
Vowel height	1.31	1.40	−2.004

**Indicates $p < 0.01$; ***indicates $p < 0.001$ (test significance is adjusted for multiple comparisons).

Tests of diagnosticity. In order to assess the extent to which combining cues resulted in accurate classification, as a test of diagnosticity we performed a multivariate linear discriminant analysis of open/closed class category. Discriminant analysis provides a classification of items into categories based on a set of independent variables. The chosen classification maximises the correct classification of all members of the predicted groups. The baseline for classifying into two classes is thus 50%, and correct classifications above 50% overall mean that there is useful information in the cues. It is possible for classification of a particular group to be below 50% and the discriminant analysis still be significant overall if the classification of the other group is much greater than 50%. Therefore, the success of the cues in the discriminant analysis is to be determined by the overall correct classification.

The open and closed class groups were weighted equally, so that the greater number of open class words did not influence the discriminant analysis to predict a greater number of items as open class. We assessed the classification using leave-one-out cross-validation. This constructs the classification on all words except for one, and assesses whether the classification correctly applies to the word that was left out. When all cues were entered simultaneously, 100% of open class and 52.7% of closed class words were correctly classified. Overall, 76.4% of words in weighted groups were classified correctly, and 98.5% of words when groups were unweighted, which was highly significant (Wilks $\lambda = 0.459$, $\chi^2 = 3667.256$, $p < 0.001$). When cues were entered in a stepwise analysis, so that only those cues that contributed significantly towards correct classification were employed, the results were identical, with 100% of open class words and 52.7% of closed class words correctly classified (Wilks $\lambda = 0.460$, $\chi^2 = 3654.594$, $p < 0.001$). Cues were entered in the following order (the first-entered cues contributed most to correct classification): presence of stress, stress position, onset complexity, syllabic complexity, reduced first vowel, and initial /ð/. Interestingly, reduced first vowel contributed even though this variable was not significantly different in terms of its mean. This suggests that it combines with the other variables towards distinguishing open from closed class words.

Most of the closed class words are high frequency, and occur in the most frequent 1000 words of the CHILDES corpus (93 of the 146), but it is still important to test that the distinctions that can be made between open and closed class words for all 5000 words apply to the more frequent set—the set of words that, children are most likely to acquire first. Kauschke and Hofmeister (2002), for instance, reported that high-frequency relational words (e.g. *up*, *down*, *on*, *under*) are acquired earliest in German, followed by high-frequency nouns and verbs. It is possible, for instance, that the majority of correctly classified closed class words are lower frequency. To test this, we performed a discriminant analysis on the 1000 most frequent words in the CHILDES corpus. When all cues were entered simultaneously, the classification was more accurate than that for the set of 5000 words. For the stepwise analysis, 100% of open class and 68.8% of closed class words were correctly classified (weighted overall 84.4% correct, unweighted overall 96.6% correct, Wilks $\lambda = 0.299$, $\chi^2 = 1017.360$, $p < 0.001$). The cues entered were: presence of stress, onset complexity, syllable length, and stress position. When all cues were entered, performance was very similar.

There is considerable power in the 16 phonological cues to discriminate between open and closed class words, and this is true both for the most frequent set of 1000 words and

the whole set of 5000 words. The next study assesses the extent to which the cues may provide information about grammatical category distinctions *within* the open class words, in particular that between nouns and verbs.

4.3. Results: distinguishing nouns and verbs

Tests of significance. We performed significance tests for each of the 16 phonological cues on measurements of the difference between the means of the 2751 nouns and the 1139 verbs in the most frequent 5000 words in the CHILDES corpus. Results of Mann–Whitney *U*-tests are shown in Table 3. Several cues were highly significantly different in terms of their means for nouns and verbs, as anticipated by previous studies. At the word level, nouns had more syllables than verbs. At the syllable level, verbs had greater onset complexity and syllabic complexity than nouns, whereas nouns had more reduced syllables than verbs. Also, verbs ended in *-ed* more often than nouns. For phoneme level cues, nouns had more coronal consonants than verbs, but fewer nasal consonants. Vowels in nouns were further back and higher than vowels in verbs. There were no significant differences in terms of stress. In Kelly and Bock's (1988) analysis, they considered only bisyllabic nouns and verbs to highlight differences in position of stress. When the whole corpus of child-directed speech was consulted, as in our analyses, such differences were not apparent. Surprisingly, though, several cues that were posited as distinguishing open and closed class words were differently distributed for nouns and verbs. Reduced syllables, for example, were found to a greater extent in nouns than in verbs.

Tests of diagnosticity. The significant differences between means indicate that several phonological cues may contribute towards classification of nouns and verbs. Yet how successful are the phonological cues in diagnostic tests for discriminating between nouns and verbs? As Shi et al. (1998) noted, grammatical categories may differ in terms of their

Table 3
Comparisons between nouns and verbs in the most frequent 5000 words for the 16 phonological cues

Phonological cue	Nouns	Verbs	*Z*
Phoneme length	4.78	4.56	−2.709
Syllable length	1.70	1.53	−6.382***
Presence of stress	–	–	–
Stress position	1.08	1.10	−2.916
Onset complexity	1.12	1.20	−3.808**
Syllabic complexity	0.66	0.67	−4.220**
Reduced syllables	0.14	0.05	−12.512***
Reduced 1st vowel	0.02	0.03	−1.148
-ed inflection	0.00	0.04	−9.275***
Coronal	0.61	0.58	−3.569*
Initial/ð/	–	–	–
Final voicing	1.19	1.23	−1.381
Nasal	0.17	0.20	−4.922**
Stressed vowel position	1.78	1.74	−1.439
Vowel position	0.76	0.64	−6.433***
Vowel height	1.36	1.20	−6.457***

*Indicates $p<0.05$; **indicates $p<0.01$; ***indicates $p<0.001$ (adjusted for multiple comparisons).

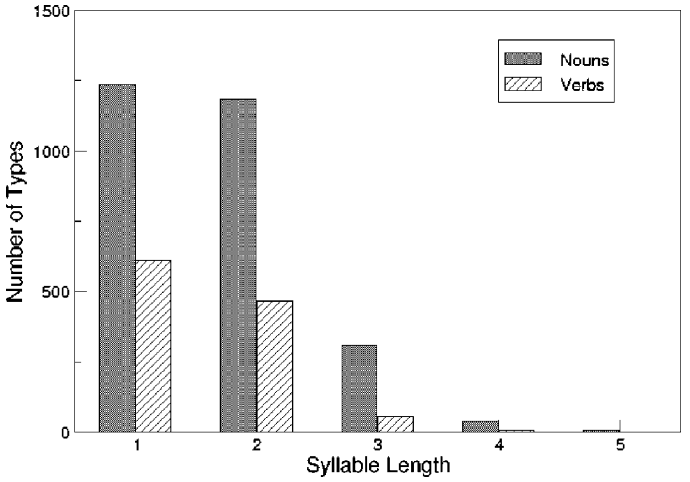

Fig. 1. Distribution of nouns and verbs for the syllable length cue.

means for a given phonological cue, but the overlap of the categories for that cue may be considerable. Figs. 1 and 2 indicate the distribution of nouns and verbs for two of the cues that reached the highest significance in the Mann–Whitney U tests. For number of syllables ($p < 0.001$), the overlap between nouns and verbs is very substantial. Applying an arbitrary cut-off at any length would result in incorrect classification of a large number of nouns or verbs. For example classifying nouns as equal to or longer than two syllables and verbs as equal to one syllable resulted in the greatest number of overall correct classifications for that cue: 55.4% of nouns and 53.5% of verbs. A Monte Carlo statistic showed that this classification was the best out of 1000 randomised analyses ($p = 0.001$), however it still meant that more than half of the nouns were incorrectly classified.

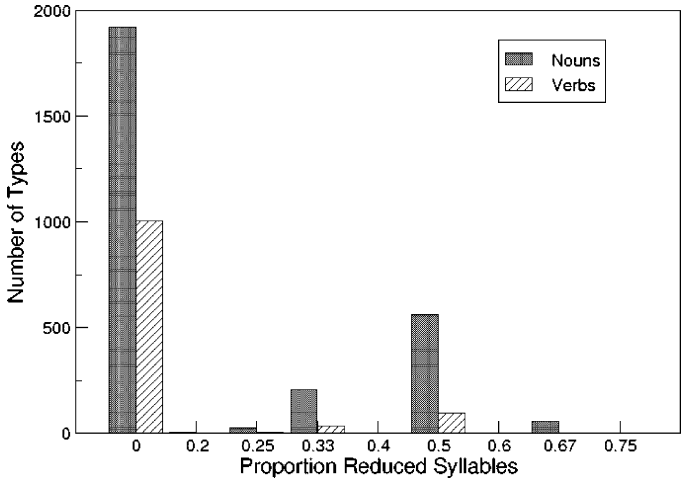

Fig. 2. Distribution of nouns and verbs for the cue proportion of reduced syllables.

For the cue that reached the highest significance-proportion of reduced syllables ($p<0.001$)—the overlap between nouns and verbs is still extremely large. The most accurate classification using a cut-off point resulted from taking nouns as having some proportion of reduced syllables and verbs having no reduced syllables, resulting in correct classification of 30.8% of nouns and 88.2% of verbs. Again, a Monte Carlo statistic indicated that the actual result was significant at $p=0.001$, though the number of incorrect classifications was very high especially for nouns.

We tested the combined contribution of all the cues towards correct classification using linear discriminant analysis, in the same way as for the open and closed class distinction above. When the phonological cues were entered in a stepwise analysis, 58.5% of nouns and 68.3% of verbs were correctly classified (63.4% overall for weighted groups, 61.3% overall for unweighted groups). This was highly significant, Wilks $\lambda=0.908$, $\chi^2=376.889$, $p<0.001$. In the stepwise analysis, cues were entered in the following order: proportion of reduced syllables, presence of -ed inflection, proportion of nasal consonants, vowel height, reduced first vowel, syllabic complexity, proportion of coronal consonants, stress position, and syllable length. Note that two cues that did not reach significance in the Mann–Whitney tests were entered in the discriminant analysis: reduced first vowel, and stress position. Results when all 16 cues were entered were similar.

Interaction with frequency. In order to explore whether correct classification was even across different frequency ranges, we divided words in the CHILDES database into groups of 1000 words (so, 1–1000th most frequent words, 1001–2000th most frequent words, 2001–3000th most frequent words, and so on). We then examined the correct classification for nouns and verbs from within these frequency groupings. For the discriminant analysis when all cues were entered, classification improved as frequency diminished. For the 1–1000th most frequent items, overall weighted correct classification was 60.1, 64.0% for the 1001–2000th group, 63.8% for the 2001–3000th group, 67.4% for the 3001–4000th group, and 67.3% for the 4001–5000th group. Results were very similar for the stepwise analysis: 60.2% for 1–1000th group, 63.2% for the 1001–2000th group, 62.9% for the 2001–3000th group, 66.0% for the 3001–4000th group, and 64.9% for the 4001–5000th group. In both analyses, performance improved for lower frequency items, but this was more marked for the analysis when all cues were entered.

The analyses we have thus far presented assess the idealised information content of the cues. Comparisons of means and discriminant analyses require that the distinctions between grammatical category are made prior to classification. However, it is more plausible that, in earlier stages of language acquisition, the child may begin to generalise from the set of acquired words to words that occur less frequently and which have not yet been acquired. In order to provide a better indication of the information sources available to the child, we assessed the usefulness of the cues for the 1000 most frequent words in the CHILDES database, and examined the extent to which a discriminant analysis function determined on these most frequent words would generalise to lower frequency groupings. For the 1000 most frequent words, the classification was less accurate than that for all 5000 words. When all cues were entered simultaneously, 48.8% of nouns and 63.3% of verbs were correctly classified in a cross-validated analysis (56.1% correct overall for weighted groups, and 53.9% correct overall for unweighted groups, Wilk's $\lambda=0.932$, $\chi^2=43.264$,

$p < 0.001$). For the stepwise analysis, 36.1% of nouns and 85.3% of verbs were correctly classified (60.7% correct overall for weighted groups, 53.4% for unweighted groups, Wilk's $\lambda = 0.948$, $\chi^2 = 32.621$, $p < 0.001$), with cues entered in the following order: proportion of reduced syllables, syllabic complexity and reduced first vowel.

We assessed the correct classification of nouns and verbs within the sets of lower-frequency words, as described above, using the discriminant analysis function derived from the most frequent 1000 words. The functions generalised well to lower-frequency items, even though these were not used to construct the function. For the function derived from all cues entered simultaneously, correct classification was higher for lower frequency words (for group 1001–2000, 60.7% correct, for group 2001–3000, 60.1% correct, for group 3001–4000, 59.5% correct, and for group 4001–5000, 64.5% correct). For the function derived from the stepwise analysis, classification was as high for lower frequency groups as the modelled data (for group 1001–2000, 60.4% correct, for group 2001–3000, 61.3% correct, for group 3001–4000, 60.7% correct, and for group 4001–5000, 62.3% correct). The functions generalise remarkably well to lower frequency items, and, when all cues were entered, the derived function classified the lower frequency items better than the data it was modelled upon.

4.4. Discussion

Several phonological cues provide some information about the open/closed class and the noun/verb grammatical category distinctions. Many of the anticipated cues for distinguishing open from closed class words were effective across a large corpus of child-directed speech in the significance tests. The results of Shi, Morgan, and Allopenna (1998) for small corpora of Mandarin and Turkish, therefore, generalise well to large-scale analyses of English. For the noun/verb distinction many of the cues that Kelly (1992) reported to be useful for distinguishing these categories were found to be significantly different in tests of significance, though with the exception of position of stress. The second-syllable stress of verbs and first-syllable stress of nouns seems to be valid when bisyllabic words only were considered. For tests of diagnosticity, when cues were combined in linear discriminant analyses, performance improved over using cues singly, with several cues contributing towards correct classification both of open/closed class words and nouns and verbs.

The tests of generalisation from distinctions formed on the most frequent 1000 words to lower frequency items suggested that phonological cues increased in their validity for low-frequency words. This was not due to the greater frequency of nouns in lower-frequency groupings, as suggested by Durieux and Gillis (2001). In our analyses, the groups were weighted to compensate for the larger noun group, so that both nouns and verbs contributed equally towards defining the discriminant function. The better performance of lower frequency items, then, is due to reasons other than greater proportion of nouns in lower frequency groupings. Rather, the low frequency words seem to correspond more closely to the phonological exemplar for their grammatical category. In Experiment 2, we test whether there is also different performance for high frequency items compared to low frequency items when the distributional cues were considered.

5. Experiment 2: testing distributional cues in grammatical categorisation

5.1. Method

Corpus preparation. The same corpus as for Experiment 1 was employed.

Cue derivation. We tested the extent to which extremely local distributional information—bigram statistics provided useful information about grammatical category. We measured the occurrence of the target word appearing after a context word in contrast to Redington et al. (1998) who assessed the two previous and the two following words, or Mintz (2003) who employed one or more preceding and following words. Our analyses, therefore, are likely to underestimate the power of distributional information in relation to previous methods. Our rationale for this was to use the context words as salient cues for the category of the next word downstream of the target word. Such context-word cues could be considered as qualitatively similar to use of phonological cues within the word: the (frequent) sound preceding the target word provides information in the same modality as that of the phonological cues of the target word. Furthermore, we anticipated improved levels of completeness over analyses that used higher-order *n*-gram information.

We made a total of 20 measures, using each of the 20 highest frequency words from the CHILDES database. The number of context words was chosen to be approximately equal to the total number of phonological cues but we ensured that these words would have occurred very frequently in the environment of the child. The 20 words were *are, no, there, this, your, that's, on, in, oh, do, is, and, I, that, what, to, a, it, the,* and *you*. Several of these words were interjections, such as *oh*, which may occur before or after any other word, yet several closed class words were in this set of 20 words which tended to delimit the class of the following word. In line with Maratsos and Chalkley's analysis, for example, *the* was one of the context words. The token frequencies of these words in the child-directed speech from CHILDES ranged from 43,022 (7913 per million) for *are* to 254,458 (46,802/million) for *you*.

In order to measure the association between the context word and the target word, we assessed the co-occurrence of context and target word bigrams using an adapted version of Dunning's (1993) log-likelihood test. This test was devised to provide an assessment of the independence of two words in a corpus. If two words *A* and *B* occur independently within a text then the probability of the occurrence of the bigram *AB* should be the same as the probability of *A* multiplied by the probability of *B*, and the log-likelihood test assesses the extent to which this is so. The test is similar to a χ^2 measure, except that it performs better when the counts of co-occurrence are taken from a small corpus.

However, in order to assess the strength of association between two words, we needed to discriminate between surprise—occurring more frequently than expected by chance—and coincidence—occurring less frequently than expected by chance. The log-likelihood test will produce a highly significant score if the two words occur more *or* less than would be expected by chance. In order to discriminate surprise from coincidence, we added an additional test of the difference between the expected frequency of the bigram *AB* and the actual frequency. We computed the difference between the independent probabilities of *A* and *B* and the probability of the bigram *AB*: $p(A)p(B) - p(AB)$. If this difference is positive,

so that actual frequency falls short of expected frequency, then we multiplied the log-likelihood test score by −1.

The resulting analysis of the adult speech from CHILDES supported the notional distinction between bigrams that were significant because they occurred more than by chance and those that occurred less than by chance. The ten most strongly associated bigrams and the 10 most strongly dissociated bigrams with the context word as the first word and target word as the second word in the bigram are shown in Table 4. Most of these low-scoring bigrams in Table 4 are composed of two high frequency words which occur together significantly less than by chance and so can be encoded as such. In effect, the context word provides negative information about the word's occurrence. The ten highest scoring bigrams also tend to be composed of two frequent words, though with a frequency of co-occurrence higher than expected by chance. The 100 highest scoring bigrams indicated a clear emerging pattern that shows a target word occurring almost exclusively with a particular context word, for example, *the beach* scores 2918.5 and occurs 443 times, whereas *beach* occurs only 57 times after any other word. The lowest scoring 100 words indicated a tendency for higher frequency pairs of words occurring much less than expected, and many ungrammatical pairings occur. For example, *a with* scores −881.5, and occurs three times, whereas *a* occurs before words other than *with* 121,549 times, and *with* occurs after words other than *a* 25415 times. Bigrams that are neither strongly associated nor dissociated in the corpus will result in a log-likelihood test score around 0. Low frequency words are such examples, as they will not occur more or less than expected by chance until they have occurred several times. Hence, a single occurrence of a word

Table 4
Top 10 and bottom 10 bigrams by association using the signed log-likelihood test from the CHILDES corpus

$-2 \log \lambda$	$k(AB)$	$k(A-B)$	$k(-AB)$	$k(-A-B)$	A	B
126256.0	32,404	222,054	34,913	6,517,058	Do	You
84203.3	18,117	10,131	236,341	6,541,840	You	Want
76595.8	20,217	234,241	22,805	6,529,166	Are	You
68140.3	18,717	151,270	40,528	6,595,914	In	The
62491.7	12,557	26,883	74,853	6,692,136	I	Don't
58961.0	10,139	12,251	77,271	6,706,768	I	Think
42924.3	11,755	14,338	242,703	6,537,633	You	Know
39968.0	11,389	77,203	57,434	6,660,403	Is	That
39559.3	13,264	156,723	42,843	6,593,599	On	The
37064.2	7260	27,364	43,291	6,728,514	That's	Right
…	…	…	…	…	…	…
−5757.6	42	169,945	121,510	6,514,932	A	The
−6014.2	393	169,594	169,594	6,466,848	The	The
−6242.1	43	254,415	87,367	6,464,604	I	You
−6595.7	3	130,067	169,984	6,506,375	The	It
−7455.3	345	254,113	129,725	6,422,246	It	You
−8557.4	77	254,381	121,475	6,430,496	A	You
−9512.0	43	130,027	254,415	6,421,944	You	It
−9712.4	463	169,524	253,995	638,2447	You	The
−12714.6	32	254,426	169,955	6,382,016	The	You
−15185.9	596	253,862	253,862	6,298,109	You	You

co-occurring with a context word will results in a log-likelihood test score close to zero, and subsequent occurrences will increase this value, but only if the co-occurrence of that word is constrained (as in *the beach*). As frequency reduces, therefore, the reliability of the association falls, and so does the log-likelihood test score. We predict that certain context-words will produce high positive scores for particular syntactic categories and high negative scores for other syntactic categories (e.g. *the* for nouns).

Our measure of co-occurrence assumes that exposure to many instances of the target word adds to the reliability of the contextual cues used for encoding that word. There is equally the possibility that the child can learn from single exposures to a word—one-shot learning. However, taking a perspective on the learning situation that the child is constructing a model of the language that minimises description length (Chater & Vitányi, 2003) requires that sufficient exposure to a stimulus is required before the co-occurrence information is used in hypothesising the distribution of a word. Onnis, Roberts, and Chater (2002) report a minimum description length algorithm that encoded bigrams as present or absent in the language in order to minimise the description length of the model of the language and the ability of the model to fit the data. However, directly applying this algorithm to encoding bigrams in child-directed speech requires a number of assumptions to be made about the parameters of the fit. The signed log-likelihood test performs a similar operation of encoding presence or absence of bigrams but does not require parameters to be set.

A further caveat to the signed log-likelihood analysis is that it ignores most of the structure of the language entirely, and in particular the co-occurrence statistics here are independent from similar co-occurrences of the target word with other context words. For example, *a* and *the* occur approximately interchangeably, and many target words that are nouns occur both with *a* and *the* in equal measure. An intelligent encoding of contextual information would pick up on this similarity, yet our analyses do not. We are therefore assuming that the starting point to the encoding of the language is the weakest hypothesis about the structure of the language, namely that every word occurs independently, and any similarities or interdependencies are ignored in the computing of co-occurrence statistics. This would seem to be consistent with the item-based learning approach suggested by Tomasello (2000).

5.2. Results: open and closed class words

We classified open and closed class words using the same CELEX categories as in Study 1A.

Tests of significance. We measured the differences in means for the log-likelihood test scores for the open and closed class words to test significance for each of the 20 context words. The results are shown in Table 5.

Tests of diagnosticity. The combined contribution of the distributional cues was assessed using discriminant analysis, precisely as for the phonological cue analyses. For the stepwise analysis, 99.6% of open class and 15.1% of closed class words were correctly classified (overall weighted correct 57.4%, unweighted correct 97.0%, Wilk's $\lambda = 0.944$, $\chi^2 = 269.983$, $p < 0.001$). Distributional cues that were entered were co-occurrence with *are, no, you, that's, on, do, is, I, to*, and *a*. Results when all cues were entered were similar.

Table 5
Comparisons between open and closed class words in the 5000 most frequent words in CHILDES for log-likelihood scores of co-occurrence with the 20 context words

Context word	Open class	Closed class	Z
Are	−4.151	643.961	−7.619***
No	−3.042	−40.100	−9.936***
There	−3.954	−5.241	−2.449*
This	14.699	−76.384	−9.309***
Your	31.658	−122.841	−14.981***
That's	10.424	216.342	−5.067***
On	−1.373	341.017	−1.064
In	2.141	567.063	−92.179*
Oh	2.904	−60.281	−11.643***
Do	−8.019	965.942	−3.769***
Is	−5.305	642.311	−3.277***
And	5.903	41.674	−2.446*
I	34.459	−285.138	−13.692***
That	7.441	−38.167	−6.899***
What	32.131	−51.992	−10.337***
To	39.435	−117.922	−7.915***
A	67.380	−372.296	−15.539***
It	5.412	−137.438	−0.992
The	86.072	−286.491	−15.295***
You	61.196	−467.532	−13.28***

*Indicates $p<0.05$, *** indicates $p<0.001$.

For the most frequent 1000 words, when all cues were entered, performance was slightly improved, with 97.2% of open class and 33.3% of closed class words correctly classified (overall weighted correct 65.3%, overall unweighted correct 90.2%, Wilk's $\lambda=0.887$, $\chi^2=99.941$, $p<0.001$). For the stepwise analysis on the most frequent 1000 words, 98.4% of open class and 24.7% of closed class words were correctly classified (61.6% weighted correct, 90.3% unweighted correct, Wilk's $\lambda=0.906$, $\chi^2=83.239$, $p<0.001$). Co-occurrence cues that were entered were *are, your, on, do, is, to,* and *a*.

5.3. Results: nouns and verbs

The same sets of nouns and verbs as employed in Study 1B were used.

Tests of significance. Table 6 shows the comparisons of the means of log-likelihood scores for nouns and verbs for every context word, for the tests of significance. Several context words were more closely related to nouns than verbs: *no, there, this, your, that's, on, in, oh, do, is, that, a,* and *the*. The context words *and, I, what, to, it,* and *you* were more closely related to verbs than nouns.

Tests of diagnosticity. For the 5000 most frequent words from CHILDES, for the stepwise analysis, 93.7% of nouns and 31.1% of verbs were correctly classified (62.4% weighted correct, 75.5% unweighted correct, Wilk's $\lambda=0.935$, $\chi^2=263.839$, $p<0.001$), with the following cues entered: *your, that's, on, do, is, and, I, to, a, it, the*, and *you*. When all cues were entered the results were very similar.

Table 6
Comparisons between nouns and verbs in the 5000 most frequent words in CHILDES for log-likelihood scores of co-occurrence with the 20 context words

Context word	Nouns	Verbs	*Z*
Are	−2.456	−9.080	−0.824
No	−0.305	−8.790	−10.342***
There	−2.970	−4.582	−3.018**
This	18.617	15.178	−16.062***
Your	56.758	−10.066	−27.280***
That's	−2.658	−11.146	−4.076***
On	2.045	−11.959	−6.922***
In	0.255	−15.134	−7.653***
Oh	−2.651	−7.105	−4.832***
Do	−4.080	−14.555	−5.387***
Is	−3.907	−11.828	−3.393***
And	1.230	5.628	−9.906***
I	−5.698	161.876	−22.608***
That	10.327	5.670	−12.385***
What	6.656	106.607	−5.739***
To	5.292	158.317	−6.942***
A	83.991	−19.824	−23.955***
It	−7.967	24.046	−10.211***
The	154.382	−37.709	−36.769***
You	−16.391	300.621	−34.44***

**Indicates $p<0.01$; ***indicates $p<0.001$.

When the correct classification was determined on the different frequency groupings, we found that correct classification was very accurate for the high frequency group, but dropped away for lower-frequencies. For the stepwise analysis, 85.8% correct for the highest frequency group, 69.6% for the 1001–2000th group, 53.5% for the 2001–3000th group, 52.4% for the 3001–400th group, and 50.4% correct for the lowest frequency group. When all cues were entered there were very similar results.

When the discriminant analysis was determined on the most frequent 1000 words, and generalisation to lower frequency groups was assessed, a similar pattern emerged. When all cues were entered, performance was good for the 1000 most frequent words (79.9% weighted correct, 81.8% unweighted correct, Wilk's $\lambda=0.735$, $\chi^2=187.075$, $p<0.001$), but close to 50% for lower frequency groupings. Similar results were found for the stepwise analysis (82.9% weighted correct for the most frequent 1000 words, 84.2% unweighted correct, Wilk's $\lambda=0.748$, $\chi^2=177.744$, $p<0.001$), with performance again around 50% for lower frequency groupings. Interestingly, the performance on the most frequent 1000 words was *better* when the discriminant analysis function was derived from all 5000 words. That is, when the function was generated to give the best fit to only the 1000 most frequent words then classification was not as accurate.

5.4. Discussion

Distributional cues in the form of a measure of local co-occurrence resulted in discriminations between open and closed class words and between nouns and verbs.

In tests of significance, considered individually, several high-frequency context words were different in terms of their associations with each grammatical category distinction. Open class words were more closely associated with 12 of the 20 context words, principally articles and pronouns. Closed class words were more strongly associated with six context words, including high frequency verbs and prepositions. The stronger association of 12 context words with open class words was partly due to many of the context words themselves being closed class words, and the co-occurrence of two closed class words was rarer than occurrence of a closed class word followed by an open class word. Furthermore, the closed class words were, on average, of higher frequency and the log-likelihood test is sensitive to this: if two high frequency words do not appear together frequently then they will receive a high negative score.

The smaller number of context words associated with verbs than nouns is perhaps due to the greater range of local contexts in which verbs occur. Verbs can be preceded by a wide variety of nouns, whereas nouns tend to be more constrained by occurring after determiners. A word that can occur in many contexts will be less strongly associated with a target word, as the overall frequency of the context word influences the log-likelihood score. As anticipated, pronouns such as *I* and *you* are more likely to occur before verbs, and the log-likelihood scores for these words with verbs have higher means than with nouns. Our hypothesis that articles such as *the* and *a* would co-occur with nouns with greater log-likelihood scores than with verbs was also confirmed. Surprisingly, words such as *oh* and *no* co-occur with greater likelihood with nouns than verbs. This runs counter to expectations that interjections can occur in any context, however, they are perhaps more like punctuation marks, and punctuation is constrained in its co-occurrence—verbs are unlikely to follow a stop or a comma, for example, except in imperative form.

When cues were combined in diagnosticity tests, open/closed class words and nouns/verbs were distinguished with a high degree of accuracy. The discriminant analyses provide a composite measure of accuracy and completeness, as the discrimination is made into only two classes, whereas Mintz's (2003) analysis, for instance, measured hits and misses for a number of frames individually. However, high accuracy and completeness resulted from combining cues. For the combined analyses, different cues were found to be useful as contributing towards successful classification to those that differed in terms of the mean of the log-likelihood scores between groupings. For high frequency items, distributional information was found to be particularly useful, but, unlike for the phonological analyses, discriminations for lower frequency nouns and verbs had lower levels of accuracy.

We have shown that bigram co-occurrences of words with a small set of high frequency words provides valuable information that can contribute towards classifying words into different grammatical categories. Furthermore, this information is most useful for high frequency words, and classifies lower frequency words with diminishing accuracy. We wanted to test whether this particular aspect of distributional information was particular to the type of distributional analysis we were considering. We therefore performed an assessment of distributional information that was different from the analysis presented above in two ways. To test whether the supervised nature of the analysis was key to the results we used an unsupervised analysis, and to test whether bigram co-occurrences were

key, we assessed words in terms of the similarity of their representations in terms of multidimensional contextual co-occurrence vectors.

5.5. Distributional information in unsupervised analyses

We replicated the method of Redington, Chater, and Finch (1998) on the 1000 most frequent words from the CHILDES database, taking co-occurrence vectors of words with the 150 most frequent words at distances of two before, one before, one after, and two after. The resulting 600 dimensional vector for each word was then compared to that of every other word in terms of Spearman's rank correlation coefficient, and a hierarchical cluster analysis was performed on the resulting matrix of similarity measures between each word. We cut the cluster analysis at the point where informativeness was maximised, which was computed from accuracy and completeness of words of the same category occurring in the same cluster. The cut-off was made at a similarity of 0.7, resulting in overall accuracy of 0.76, and completeness of 0.41.[3] Baseline performance for a randomised grammatical classification was accuracy of 0.26 and completeness of 0.14. The overall performance of the analysis was similar to that of Redington, Chater, and Finch (1998), though they employed a cut-off of 0.8 to maximise informativeness.

We then performed the same analysis for the 4001–5000 most frequent words in the CHILDES database. We hand-coded the words that did not have grammatical categories from the CELEX database (e.g. proper nouns, interjections, numerals). With a cut-off of 0.7, accuracy of the clustering was 0.90, but completeness was very low at 0.03: the analysis produced many small clusters that had high agreement of category but were far from containing all the members of the same grammatical category. The random baseline was 0.44 accuracy and 0.01 completeness. We used the same cut-off as for the 1000 most frequent words as the arbitrary cut-off point has to be decided on the basis of all items in the space, and is likely to be more heavily influenced by the more frequent items in the language environment.

In summary, we found a similar effect of frequency to that of the original localist co-occurrence analysis, even though this assessment was very different in terms of being unsupervised and considered the closeness of words in terms of their co-occurrence with a large number of words in several different positions.

We now present analyses of combining the phonological and distributional cues for grammatical categorization.

6. Experiment 3: combining phonological and distributional cues

Fig. 3 shows the correct classification of nouns and verbs based on the discriminant analyses of phonological or distributional cues entered separately for

[3] Note that an assessment of classification comparable to that of the discriminant analyses cannot be made, as the cluster analysis produced several clusters. If the cluster produced two clusters, then completeness and accuracy would correspond to correct/incorrect classifications from the discriminant analyses.

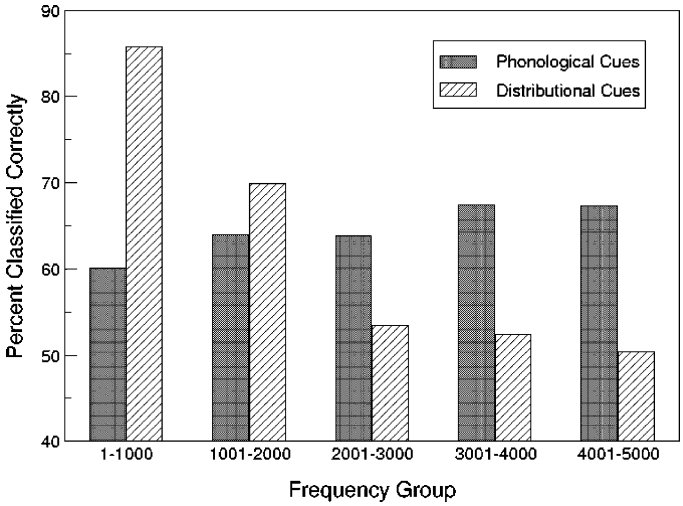

Fig. 3. The relative contribution of phonological and distributional cues for classifying nouns and verbs of different frequency groupings. For high frequency items, distributional cues result in successful classification but perform more poorly on low frequency items. In contrast, phonological cues classify lower frequency items with more accuracy than higher frequency items.

different frequency groupings. For high-frequency items, distributional information is extremely useful, but drops off dramatically for lower frequency items. For the phonological cues, the opposite pattern is observed: better performance for lower frequency words.[4] We predicted that phonological and distributional cues would contribute differentially towards correct classification. Combining the two ought to lead to more accurate classification for the high frequency items, and better generalisation to lower frequency items. We tested this for open/closed class words and for nouns and verbs.

6.1. Method

We performed the same corpus preparation and used the same grammatical categorisations as for Studies 1 and 2. We report only the tests of diagnosticity below, as tests of significance are identical to those performed on the separate analyses of phonological and distributional cues.

[4] Note that Soreno and Jongman's (1990) finding that difference in vowel position was assessed on high frequency nouns and verbs, though we found a significant difference on all 5000 words (Table 3). There is some coherence in phonological cues for the high frequency words, as indicated by better than chance discrimination on the 1000 most frequent words from CHILDES. It is possible that certain phonological cues provide better discrimination for higher frequency words, but the general picture of greater reliability of cues for lower frequency words remains true.

6.2. Results: open and closed class words

Tests of diagnosticity We performed a linear discriminant analysis on the open/closed class words distinction for the 5000 most frequent words, entering all 16 phonological cues and all 20 distributional cues. 99.9% of open class and 52.7% of closed class words were correctly classified (76.3% weighted correct, 98.5% unweighted correct, Wilk's $\lambda = 0.454$, $\chi^2 = 3703.305$, $p < 0.001$). For the stepwise analysis, performance was very slightly improved, with 100% of open class words and 52.7% of closed class words correctly classified (76.4% weighted correct, 98.5% unweighted correct, Wilk's $\lambda = 0.457$, $\chi^2 = 3690.462$, $p < 0.001$). In the stepwise analysis, six phonological cues and three distributional cues were entered in the following order: presence of stress, stress position, onset complexity, co-occurrence with *that's*, syllabic complexity, co-occurrence with *the*, reduced first vowel, initial/ð/, and co-occurrence with *there*.

6.3. Results: nouns and verbs

Tests of diagnosticity. When all phonological and distributional cues were entered into a discriminant analysis of all nouns and verbs in the most frequent 5000 words of the CHILDES corpus, 67.0% of nouns and 71.4% of verbs were correctly classified (69.2% weighted correct, 68.3% unweighted correct, Wilk's $\lambda = 0.843$, $\chi^2 = 666.923$, $p < 0.001$). When the different frequency groupings were distinguished, performance was good across the board, 79.7% for the 1–1000th group, dropping less dramatically as frequency than for the distributional cues alone to 67.4% for the 4001–5000th group.

For the stepwise analysis on the 5000 most frequent words, 10 phonological cues and 13 distributional cues were entered in the following order: reduced syllables, co-occurrence with *the*, *-ed* inflection, co-occurrence with *do*, proportion of nasals, co-occurrence with *your*, vowel height, syllabic complexity, reduced first vowel, co-occurrence with *is*, co-occurrence with *a*, co-occurrence with *that's*, proportion of coronals, stress position, syllable length, co-occurrence with *I*, co-occurrence with *and*, co-occurrence with *you*, co-occurrence with *it*, onset complexity, co-occurrence with *in*, co-occurrence with *this*, and co-occurrence with *are*. 66.3% of nouns and 71.9% of verbs were correctly classified (69.1% weighted correct, 67.9% unweighted correct, Wilk's $\lambda = 0.848$, $\chi^2 = 642.520$, $p < 0.001$). Performance for each frequency grouping dropped with frequency though was still well above chance levels for the lowest frequency group, ranging from 79.6% for the highest frequency grouping, to 68.5% for the lowest frequency grouping.

6.4. Discussion

We have presented analyses of the potential information available from phonological and bigram distributional sources for distinguishing different grammatical categories. Several phonological cues, reported in the literature, were shown to contribute towards distinguishing open from closed class words, and nouns from verbs in a large child-directed speech corpus. Previous analyses of phonological cues have assessed small child-directed speech corpora, or lexica derived from written or adult-to-adult speech.

The analyses presented above indicate that the conclusions of these previous studies apply equally to the entire CHILDES corpus, which currently stands as the best available approximation to the child's linguistic environment. The bigram distributional cues were also found to contribute towards accurate classification of open and closed class words and nouns and verbs. These analyses of distributional information were novel in that they considered only the association between the target word and a small set of high frequency context words. We made the assumption that the child quickly learns the form of these frequently occurring words, and can use them to classify the words that follow them. We have also made the assumption that information that proves useful in the environment for categorisation is used by the child in the early processes of language acquisition.

Combining cues across the phonological and distributional modalities provided better classification overall than using either type of cue alone. The contribution of cues was not always additive, some cues were not entered into the combined discriminant analysis though they were entered into the separate analyses, and other cues were only entered in the combined analyses. For the open/closed-class distinction, there were more phonological cues than distributional cues entered. In the analyses when distributional cues were considered alone, ten cues were entered for the open/closed class distinction. In the combined analysis, only three distributional cues contributed to the accuracy of the classification. Two of these cues were not entered when distributional cues were considered alone. Co-occurrence with *the* and with *there* were only useful in the combined analysis. In contrast, the same phonological cues were entered in the combined analysis as when phonological cues were considered alone for distinguishing open class and closed class words. The combined analysis performed with the same overall accuracy as when phonological cues were considered alone (76.3% correct compared to 76.4% correct, respectively). This suggests that there is overlap in the information provided by phonological and distributional cues. For the open/closed class distinction, phonological information appears to be more useful than the bigram distributional cues we have considered.

For the noun/verb distinction, the same nine phonological cues used in the phonological cue classification were entered in the combined analysis, and one additional phonological cue was employed in the combined analysis: onset complexity. Co-occurrence with *on, to*, and *it* were used in the distributional cue analysis but not entered in the combined analysis. Co-occurrence with *the, your, and, it, in*, and *this* were entered in the combined analysis but not the distributional cues alone analysis. Otherwise, the cues contributed additively in the current analysis. The classification resulting from the combined analysis was more accurate than that of the phonological cues or the distributional cues alone (which achieved accurate weighted classification of 64.3 and 62.4%, respectively). This seems to be due to the better classification for high frequency items due to the distributional cues and better classification for the lower frequency items due to the phonological cues.

Phonological and distributional cues are therefore not orthogonal, and combinations of cues may over-ride the contribution of other cues, for example in the combined analysis of nouns and verbs fewer distributional cues were entered into the analysis as a result of phonological cues providing the discrimination that distributional cues would perform in the absence of phonological information.

We now provide an experimental test of the availability of these sources of information in language learning. We trained adults to learn an artificial language that varies the richness of distributional and phonological cues and we assessed the effectiveness of learning distinct categories under this variation. In particular, we tested the prediction from the corpus analyses that words with rich distributional information will be learned more easily than those with impoverished distributional cues. Similarly, words which are coherent with regard to phonological cues of the same category will be learned more easily than those without this coherence, but this will be most emphatic for words which are low-frequency and hence with poor distributional information. Thus, we predict main effects of richness of distributional cues, richness of phonological cues, and an interaction between the cue types.

7. Experiment 4: artificial language learning of bigrams

We adapted Valian and Coulson's (1988) artificial language such that category words were presented with different frequencies during training. Our hypothesis was that the association with marker-words would be learned more quickly for the high-frequency category words than the low-frequency category words. We also varied the extent to which there was coherence within the two categories of words. All the words within a category either shared several phonological properties, or none. We were also interested in whether this pattern changed across learning. Is phonological information particularly useful in the early stages of learning, or does it present with a stable influence across time? To this end, we tested participants twice on their acquisition of the language structure.

7.1. Method

Subjects. Twenty-four undergraduate students at York University participated in the study for course credit. All participants were first language English speakers.

Grammar. As in Valian and Coulson's (1988) study, sentences contained four words, made up of two phrases: *aA* and *bB*, where *a* and *b* were marker-words and *A* and *B* were category words, selected from sets of 6. Sentences were of the form *aAbB* or *bBaA*. Three words from each of the A and B categories occurred twice as frequently as the other three words in the same category, thus each category contained both high and low frequency words.

Stimuli. Eighteen training sentences were constructed such that the high-frequency category words occurred four times each and the low-frequency category words occurred twice each. There were an equal number of *aAbB* and *bBaA* sentences. Each category word occurred an equal number of times in the first and the second phrase in the sentences, and no two category words occurred in the same sentence more than once.

There were two distinct sets of test sentences, each comprised of 12 sentences that were compatible with the language structure, but had not occurred during the training phase. 6 sentences were comprised of two high-frequency category words, and the other 6 sentences contained two low-frequency category words. Two distinct sets of 12 sentences that were incompatible with the language structure were also included. In each

set, 6 of the incompatible sentences violated the link between one marker-word and one category word (e.g. *aAaB*), termed by Valian and Coulson a Type 3 error, and 6 violated the link between both marker-words and the category words (e.g. *aBbA*), a Type 4 error. For the Type 4 sentences three were composed of two high frequency category words, and the other three contained two low frequency category words. For the Type 3 sentences, two contained two high frequency category words, and two contained both low frequency category words. It was not possible to counterbalance the frequency and position of occurrence of each category word in the test phase without having two Type 3 sentences with one high frequency and one low frequency word. In one case, the marker-category word violation was with the high frequency word, and in the other case the violation was with the low frequency word. We grouped the high-frequency violation with the other high frequency Type 3 sentences, and did the same for the low-frequency violation sentence. In each test set, each category word occurred four times—twice in a compatible sentence and once in an incompatible sentence. For each test set, there were thus six high-frequency and six low-frequency compatible sentences, and six high-frequency and six low-frequency incompatible sentences. Valian and Coulson (1988) included other types of incompatible sentences, where the order of marker words and category-words was altered (e.g. *aABb*). However, their participants quickly learned to reject these sentences, and so we omitted them from our testing.

The marker-words were *alt* and *erd*, as in Valian and Coulson's (1988) study. For half the participants *alt* marked the *A* category and *erd* marked the *B* category, and for the other participants this was reversed (we refer to these as dialects 1 and 2). The category words were all monosyllabic nonsense words. In the phonologically coherent condition, *A* and *B* category words shared phonological properties which were found to distinguish different lexical categories in our corpus analyses. One set of words had consonant clusters at the onset and nucleus, had rounded, low vowels, and contained nasals and stops but no fricatives. The other set of words contained no consonant clusters, had unrounded, high vowels, contained no nasals or stops, but only fricatives. The first set was: *blint, dreng, gwemb, klimp, prienk*, and *tweand*, and the second set was: *foth, shufe, suwch, thorsh, vawse*, and *zodge*. Each word overlapped no more than two phonemes with another word in the same set, but did not overlap at all with words in the other set. Three words in each set were high-frequency and three were low-frequency. In the incoherent condition the three high-frequency words from set 1 were exchanged with the three low frequency words in set 2. The training sentences for the phonologically coherent condition are shown in Appendix A.

Procedure. The experiment was administered on a computer, and participants were tested individually in a quiet room. Participants were instructed that they would see sentences in a nonsense language containing meaningless words, and that they were to learn all they could about the patterns of the language. Participants were then shown a list of the vocabulary items and asked to read them aloud. When they had done this, they pressed a key on a computer keyboard and the training sentences began. Sentences were presented at the centre of the screen in 18 point bold Courier font. Sentences appeared for 10 seconds and participants were asked to read them aloud. Two random permutations of the 18 training sentences were presented, and then the participant was informed that the testing phase would begin.

In the testing phase, the participant was requested to read aloud the test sentence and press the *y* key on the keyboard if the test sentence was compatible with the pattern of the language, and the *n* key if the test sentence was incompatible. Participants were informed that half the test items were compatible and half were incompatible with the language. Each test sentence remained on the screen until the participant made their response at which point the next test item appeared. The training phase was then repeated, followed by another test phase using the second set of test sentences. After the second test phase was completed the participant was asked to sort 12 cards, each showing one of the category words used in the study, into two groups depending on which words the participant thought went together. The cards were shuffled after each participant.

7.2. Results

In the card-sorting task, two participants sorted the words into 6 groups of two sentences. These participants were judged to have misunderstood the task and so were omitted from the analyses. This left 12 participants in the phonologically coherent condition, and 10 participants in the incoherent condition. We scored the extent to which words from the *A* category were grouped together by the participant. From a maximum score of 6, the phonologically coherent condition correctly grouped 5.17 cards, and the incoherent group grouped 3.90 cards, which was significantly less, $t(20) = 2.917$, $p < 0.01$. Groupings in the phonologically coherent group differed significantly from chance level of 3.91 cards correctly sorted, $t(11) = 4.242$, $p < 0.001$, but was at chance for the incoherent condition, $t(9) = -0.016$, $p = 0.988$. This indicated that independent categorisation of the sets of words was successfully achieved in the coherent condition.

We scored the number correct for high and low frequency items for compatible sentences, and number correctly rejected for each type of incompatible sentence (*aAaB* and *bAaB*) at each test phase. The mean correct responses for the coherent/incoherent condition for each sentence type are shown in Table 7. From Table 7 it can be seen that

Table 7
Performance on compatible and incompatible sentences in Experiment 4, for testing time 1 and 2, distinguished by phonologically coherent and incoherent groups

Phonological condition	High frequency			Low frequency		
	Compatible	Type 3	Type 4	Compatible	Type 3	Type 4
Time 1						
Coherent	5.50 (0.67)	1.67 (0.98)	1.75 (1.06)	4.83 (1.19)	1.33 (1.15)	2.25 (0.87)
Incoherent	4.80 (0.92)	0.90 (0.88)	1.60 (1.17)	4.30 (1.06)	0.50 (0.85)	1.00 (0.67)
Time 2						
Coherent	5.50 (0.67)	1.75 (1.14)	1.92 (1.38)	4.92 (1.73)	1.92 (1.00)	2.17 (1.03)
Incoherent	4.80 (0.92)	1.50 (1.08)	1.70 (1.06)	3.60 (1.51)	1.60 (0.70)	1.40 (0.97)

Numbers indicate mean correctly accepted or rejected, with standard deviation in parentheses. Type 3 incompatible sentences violated the order of marker-words (e.g. *aAaB*), Type 4 incompatible sentences violated couplings of marker-words and category-words (e.g. *aAbA*). For compatible sentences scores are from a maximum of 6, for incompatible sentences scores are out of 3.

poorer performance results for all word types in the phonologically incoherent condition. For frequency, participants tended to accept sentences containing high frequency category words and reject sentences containing low frequency words. Hence, there was a tendency towards higher correct responses to low frequency Type 3 and 4 sentences. To account for this bias, the sum of the correctly accepted and correctly rejected responses was taken as the independent variable in our analyses.

We performed an ANOVA on number of correct responses with phonological coherence as a between-subjects variable, and frequency (high/low) and time (first/second test) as within-subjects variables. An ANOVA that also included dialect as a between-subjects variable did not result in significant effects involving dialect, and so we report the simpler, more powerful design that omits dialect. There was a significant main effect of coherence, with higher scores in the coherent condition than the incoherent condition (8.88 and 6.93, respectively, from a maximum 12), $F(1, 20) = 7.04$, $p < 0.05$. There was also a significant main effect of frequency, with higher scores for the higher-frequency sentences (8.35 and 7.45, respectively), $F(1, 20) = 13.15$, $p < 0.005$. There was no significant effect of time, $F(1, 20) = 2.06$, $p = 0.17$. As predicted, there was a significant interaction between frequency and coherence, $F(1, 20) = 5.15$, $p < 0.05$, with coherence making a greater impact for the lower-frequency sentences. The interaction is shown in Fig. 4. No other interactions were significant (all $F < 1$). Post hoc comparisons indicated that, for the high frequency words, there was no significant difference between scores for the phonologically coherent and incoherent groups, $t(20) = 1.74$, $p = 0.19$, but there was a significant difference for the low frequency words, $t(20) = 3.35$, $p < 0.01$.

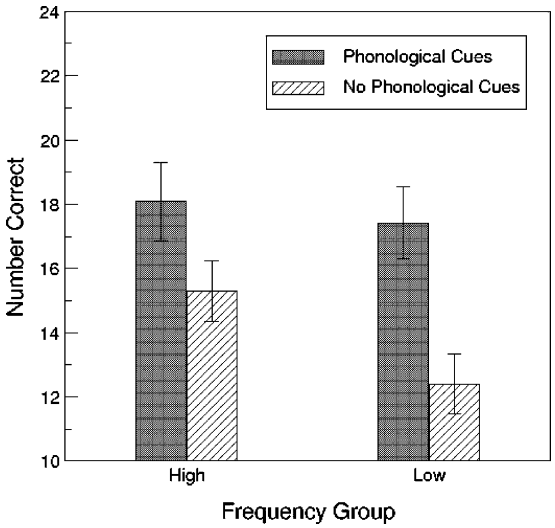

Fig. 4. Number of correct responses for sentences containing high or low frequency category words for the phonologically coherent and incoherent conditions in Experiment 4. Results are combined from first and second tests, and are out of a maximum 24.

Type 3 error sentences (of the form *aAaB*) could be correctly rejected on the basis of the participant learning that repetition of the anchor words was illegal, rather than learning the *A*/*B* category distinction. Hence, better than chance performance may be due to learning to reject this repetition rather than learning word categories. To test this we performed an ANOVA on the combined score of correct sentences and Type 4 error sentences (of the form *aAbA*), and omitted scores on Type 3 error sentences. As in the previous analysis, phonological coherence, frequency and time were factors. The results were significant in the same way as the original analysis. There were significant main effects of coherence, $F(1, 20) = 8.10$, $p < 0.05$, and frequency, $F(1, 20) = 17.65$, $p < 0.001$, and a significant interaction between coherence and frequency, $F(1, 20) = 8.10$, $p < 0.05$. All other main effects and interactions were not significant (all $F < 1$).

7.3. Discussion

The results indicated that people could learn to categorise words using bigram distributional information. In the card sorting task at the end of training, the participants scored better than chance only in the phonologically coherent group. Assessing which words people were better at categorizing was our next step, and assessment of performance on the testing sentences indicated that the higher frequency words were learned with greater ease. For the phonologically incoherent group, performance was still less accurate after the second training stage for the low frequency words than was performance on the high frequency words after the first training stage. Further, the interaction between phonological and distributional information seemed to be consistent over time: there was a similar interaction between these factors at both stages of testing.

Distributional information was equated with frequency in this study as more occurrences of a word enriches the contextual information for that word. In the artificial language, as in the corpus studies for words with constrained contexts, the log-likelihood test scores for words increase as more instances of the word are experienced. Consequently, we have shown that distributional information is less reliable and cannot be used so effectively for low frequency words in the artificial language as for the high frequency words.

The use of phonological cues follows a different pattern. For the low frequency words, phonological coherence resulted in scores that were over 20% more accurate than those for the phonologically incoherent group. There was a small, but non-significant difference between the scores for the high frequency words, indicating that phonological cues are particularly useful for classifying low frequency words. Participants were not informed that there were phonological cues in the stimuli, but their performance was nevertheless influenced by the presence of these cues. Participants managed to correlate particular cues with certain distributional patterns of words.

The principal results of the artificial language learning experiment were successful. However, there were limitations to the study with regard to issues of categorization in language acquisition. First, the artificial language was idealized in that distributional information was a perfect cue for word category. This is unlike the situation for language learning where distributional information is noisy and not always reliable.

Yet the overall pattern of enriched distributional information for higher frequency words was shared across the corpus analyses and the language learning experiment. Similarly for the phonological cues, in the coherent condition the cues matched the word category perfectly, whereas in the corpus analyses phonological cues were unreliable particularly for high frequency words. A better match to the corpus data is found in comparing the high frequency words in the phonologically incoherent condition with the low frequency words in the phonologically coherent condition. This matches the corpus data in that high frequency words have less reliable phonological cues but better distributional information, and low frequency words have more reliable phonological cues but worse distributional information. For both these cases, performance is better than the baseline of impoverished distributional and phonological information indicated by performance on the low frequency words in the phonologically incoherent condition.

8. General discussion

When cues were considered jointly in the discriminant analyses, classification accuracy increased over when single cues were considered. Cues provided additive value in the classification, contributing towards classification of different items. This was especially true when phonological and distributional cues were considered together. We found confirmation for our hypothesis that phonological and distributional information contributed differentially towards categorisation. At points where distributional information was better for classification—the high-frequency items—phonological cues were found to be of less value. Conversely, for the lower-frequency items, where distributional information was less useful, phonological information contributed towards more accurate classification. It is possible that the absence of quality distributional information for lower-frequency items pressures the phonological forms to remain distinct and true to the grammatical categories, at least along the general lines illustrated by the discriminant analyses. Because there is no other information available for these items, phonological information becomes particularly useful, and therefore preservation of any such information to assist in categorisation would be encouraged. However, this interaction between effectiveness of cues may be more subtle than just interacting with frequency. For example, Christiansen and Monaghan (in press) found that verbs are better classified with phonological information than distributional information, whereas nouns demonstrate the reverse pattern.

Another point at which there is a complementary contribution of phonological and distributional information is with respect to the open/closed class distinction. Classifications based on distributional cues were less accurate for this distinction than for the noun/verb distinction, even for the highest-frequency grouping. Phonological information provided more information about category for open/closed class words than was apparent for the high-frequency noun/verb distinction. Bigram distributional cues were not so useful for determining closed class category as the co-occurrence of closed class words with the 20 most frequent context items was rare. The compensation of phonological information for determining the open/closed class distinction is consonant

with the discussions in (Shi, Morgan, and Allopena, 1998) on the added value provided by such phonological cues for discriminating function words from content words in fast, online language processing tasks (Shillcock, Kelly, & Monaghan, 1997).

This provides a potential explanation for why single, highly unreliable cues seem to be so powerful in determining categorisation of nonwords (Cassidy & Kelly, 1991; 2001). Fig. 1 indicates the paucity of the syllable length cue for accurately distinguishing nouns from verbs, and yet this cue alone appeared to determine whether a nonword was used in a noun context or a verb context. Phonological cues prove to be more useful and reliable for lower frequency items. Their use compensates for the absence of distributional information for words that occur rarely. Nonwords are low frequency words *par excellence*, with the participant never before coming across the stimulus, and the nonword is therefore presented without any distributional information at all. In effect, the phonological information is the only clue for category. A caveat is necessary, however, as the particular nonword materials used by Cassidy and Kelly in both their studies differ on a number of other phonological cue dimensions from the list of 16 we have considered in this paper. Nevertheless, the point remains that phonological information is more reliable for low frequency items, and, in the absence of any other information source, provides a valuable cue to grammatical category for classification: the usefulness of the phonological cues for low frequency words validates the use of these cues under such conditions.

It is possible that there are other phonological cues that may be useful for grammatical categorisation that we have not yet considered. Phonological features at the word-level, the syllable-level, or the phoneme-level may be useful that we have not yet discovered, but are used by the child in category acquisition. The increase in correct classifications found by Durieux and Gillis (2001) when they encoded onset, nucleus and coda in terms of the phoneme clusters, rather than in terms of more generic phonological cues, suggests that there may be additional cues that have not yet been discussed in the literature. A large-scale search of correlations between categories and phonological features would be required to comprehensively determine the set of useful phonological cues. Also, we have not yet addressed acoustic cues that may be useful for distinguishing categories. Shi et al. (1998) found that pitch contours distinguished open and closed class words, for example, which may contribute additionally to the classification we have presented here. Such analyses are beyond the remit of this paper, as the large corpora we assessed were of transcribed speech.

Equally, there may be other characterisations of the potential distributional information to the child that may have resulted in different results. We based our account on very few assumptions about the child's ability to learn local associations between a small set of high frequency context words and each target word. Bigram statistics are available and used in speech segmentation (e.g. Saffran et al., 1996), and the artificial language learning experiments using *anchors* indicate that high frequency words interjecting in the language provide a useful scaffold for constructing the category of the next word (Foss & Jenkins, 1966; Valian & Coulson, 1988). We also assumed that the associative strength would be determined by the relative frequencies of each word in the pair, and their co-occurrence frequency. We showed above that an unsupervised cluster-based analysis on high- and low-frequency words based on the method of Redington et al. (1998) also performed poorly for low frequency words when the same cut-off level as that for the high-frequency

words was used. The main point we make from the distributional analyses is that distributional information provides less reliable and effective information about category for low-frequency items, and this is true for two very different analyses of distributional information in grammatical categorisation.

The results of the artificial language learning experiment indicate that, in learning the categories of nonsense words, adults utilize distributional information at the bigram level and this is more useful for high frequency words. Additionally, phonological information makes the largest difference in learning for low frequency words. Additional evidence for the learning of bigrams for categorisation comes from studies of gender categorisation. In learning the gender of words semantic properties are not relevant for the categorisation of many nouns, and therefore some combination of distributional and phonological cues are alone critical for forming this category distinction (Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Frigo & McDonald, 1998; Karmiloff-Smith, 1979; MacWhinney, Leinbach, Taraban, & McClelland, 1989; Mills, 1986). Indeed, Braine (1987) has suggested that learning bigram categories is only possible when there is a partial correlation among the categories in terms of phonological or semantic properties (p. 84). Braine et al. (1990) found that children learned the reference for nonwords better when there were correlates between phonological form and grammatical category. The results of Experiment 4 indicated that category learning can proceed without phonological cues across the whole category, though performance was much improved when the correlation was available.

To what extent do these results illuminate our understanding of the child learning their first language? A principal contribution of this paper has been to indicate the extent to which cues are available in the language environment for the child. The artificial language learning experiment indicates that adults are sensitive to this distributional and phonological information and draw on this for categorisating words in an artificial language. Infants have been shown to be sensitive to bigram distributional information, and also that they have some knowledge of phonological distinctions between words from different grammatical categories (e.g. Shi, Werker, & Morgan, 1999). But the question remains how the child learns which cues are useful and when they are applicable. For low frequency words, the co-occurrence statistics of words with high frequency, high salience, context words are not available—a log-likelihood score around 0 provides no information about category. Therefore the child has no option but to look elsewhere for information about category. The distributional analyses we have performed on child-directed speech indicate that distributional information is most useful for high frequency words, and the artificial language learning experiment indicates that phonological information is compensatory for learning of low frequency words. The child learns high-frequency, concrete words earlier, and concrete words tend to have more limited contexts than abstract words (Monaghan, Shillcock, & McDonald, 2004), and thus have more powerful distributional cues. Though weak, phonological cues cohere to a certain extent in high frequency, early-learned words, and this slight coherence makes learning easier, as illustrated in the learning experiment. The subsequent correlation of phonological cues with early learned instances of categories might then be extended to lower frequency items for which distributional information is not available. As more words are added to the child's vocabulary, the correlation between phonological cues and category increases,

until for very low frequency, or new, words, the phonological information can alone determine a decision about category membership. We have provided a detailed treatment of the information available in child-directed speech, but future work is required to elaborate the precise use of cues in language acquisition by children. Yet, the analyses we have presented here provide a framework for the potential and comparative value of information from different sources, and a foundation for theories of how such information may be used to begin the process of bootstrapping grammatical category information in language acquisition.

Acknowledgements

All three authors were supported by Human Frontiers of Science Program grant RGP0177/2001-B. The second author was also supported by European Commission Project grant number HPRN-CT-1999-00065.

Appendix A

Training sentences used in dialect 1 of the phonologically coherent condition in Experiment 4.

alt tweand erd foth
erd vawse alt tweand
alt tweand erd zodge
erd thorsh alt tweand
erd foth alt dreng
alt dreng erd suwch
alt dreng erd thorsh
erd shufe alt dreng
alt klimp erd vawse
erd suwch alt klimp
erd zodge alt klimp
alt klimp erd shufe
alt gwemb erd foth
erd vawse alt gwemb
alt prienk erd vawse
erd suwch alt prienk
erd foth alt blint
alt blint erd suwch

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1996). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, *9*, 321–324.

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

Bernstein Ratner, N., & Rooney, B. (2001). How accessible is the lexicon in Motherese?. In J. Weissenborn, & B. Höhle, *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition* (Vol. 1). Amsterdam: John Benjamins, 71–78.

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*. Hillsdale, NJ: Lawrence Erlbaum Associates, 65–87.

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.

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 thap-Constraints on the pronunciation of new written words. *Quarterly Journal of Experimental Psychology*, *33*, 375–396.

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.

Chater, N., & Vitányi, P. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*, 346–369.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Christiansen, M. H., & Dale, R. A. C. (2001). *Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates pp. 220–225.

Christiansen, M. H., & Monaghan, P. (2004). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: how children learn verbs*. Oxford: Oxford University Press, in press.

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.

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, *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1). Amsterdam: John Benjamins, 189–229.

Finch, S. P., & Chater, N. (1992). *Bootstrapping syntactic categories Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foss, D. J., & Jenkins, J. J. (1966). Mediated stimulus equivalence as a function of the number of converging stimulus items. *Journal of Experimental Psychology*, *71*, 738–745.

Fries, C. C. (1952). *The Structure of English: An Introduction to the Construction of English Sentences*. New York: Harcourt, Brace and 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.

Gleitman, L. R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner, & L. R. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge, UK: Cambridge University Press, 3–48.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Harris, Z. S. (1951). *Structural Linguistics*. Chicago: University of Chicago Press.

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

Karmiloff-Smith, A. (1979). *A functional approach to child language: A study of determiners and reference*. Cambridge, UK: Cambridge University Press.

Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language*, *29*, 735–757.

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*. Mahwah, NJ: Lawrence Erlbaum Associates, 249–262.

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 389–403.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *The Psychology of Learning and Motivation*, *7*, 1–41.

Lickley, R. J., & Bard, E. G. (1998). When can listeners detect disfluency in spontaneous speech? *Language and Speech*, *41*, 203–226.

Macnamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–14.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, *28*, 255–277.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

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, *Children's Language* (vol. 2). New York: Gardner Press, 127–214.

Marchand, H. (1969). *The Categories and Types of Present-day English Word-formation* (2nd Edition). Munich, Federal Republic of Germany: C.H. Beck'sche Verlagsbuchhandlung.

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

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., & Christiansen, M. H. (2004). *What distributional information is useful and usable in language acquisition? Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Monaghan, P., Shillcock, R. C., & McDonald, S. A. (2004). Hemispheric asymmetries in the split-fovea model of semantic processing. *Brain and Language*, *88*, 339–354.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of grammatical categories. In J. L. Morgan, & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates, 263–283.

Onnis, L., Christiansen, M. H., Chater, N., & Gómez, R. (2003). *Reduction of uncertainty in human sequential learning Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Onnis, L., Roberts, M., & Chater, N. (2002). *Simplicity: A cure for overregularizations in language acquisition Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes*, *13*, 129–191.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Schütze, H., (1993). Part-of-speech induction from scratch. *Proceedings of the 31st annual meeting of the association for computational linguistics*. Columbus, Ohio.

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.

Shillcock, R. C., Kelly, M. L., & Monaghan, P. (1997). Modelling within-category function word errors in language impairment. In W. Ziegler, & K. Deger (Eds.), *Clinical Linguistics and Phonetics*. London: Whurr.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, K. H. (1969). Learning co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, *8*, 319–321.

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

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Science*, *4*, 156–163.

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

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.

Zipf, G. K. (1935). *Psycho-Biology of Languages*. Cambridge, MA: MIT Press.