ELSEVIER

Cognitive Science 26 (2002) 393–424

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

The distributional structure of grammatical categories in speech to young children

Toben H. Mintz [a,*], Elissa L. Newport [b], Thomas G. Bever [c]

[a] *Department of Psychology, University of Southern California, SGM 501, MC 1061, Los Angeles, CA 90089-1061, USA*
[b] *Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA*
[c] *Department of Linguistics, University of Arizona, Tucson, AZ 85721, USA*

Received 28 April 2000; received in revised form 1 October 2001; accepted 18 February 2002

Abstract

We present a series of three analyses of young children's linguistic input to determine the distributional information it could plausibly offer to the process of grammatical category learning. Each analysis was conducted on four separate corpora from the CHILDES database (MacWhinney, 2000) of speech directed to children under 2;5. We show that, in accord with other findings, a distributional analysis which categorizes words based on their co-occurrence patterns with surrounding words successfully categorizes the majority of nouns and verbs. In Analyses 2 and 3, we attempt to make our analyses more closely relevant to natural language acquisition by adopting more realistic assumptions about how young children represent their input. In Analysis 2, we limit the distributional context by imposing phrase structure boundaries, and find that categorization improves even beyond that obtained from less limited contexts. In Analysis 3, we reduce the representation of input elements which young children might not fully process and we find that categorization is not adversely affected: Although noun categorization is worse than in Analyses 1 and 2, it is still good; and verb categorization actually improves. Overall, successful categorization of nouns and verbs is maintained across all analyses. These results provide promising support for theories of grammatical category formation involving distributional analysis, as long as these analyses are combined with appropriate assumptions about the child learner's computational biases and capabilities.

Keywords: Distributional structure; Grammatical categories; Young children

* Corresponding author. Tel.: +1-213-740-2253.
E-mail address: tmintz@usc.edu (T.H. Mintz).

0364-0213/02/$ – see front matter
PII: S0364-0213(02)00072-1

1. Introduction

Two primary factors determine the outcome of language acquisition: the language learner's linguistic environment, and the properties of her computational and representational systems. Aspects of acquisition differ in the degree to which they are driven by the input. For instance, the learner's vocabulary is strongly shaped by the particular words to which she is exposed. At the same time, many aspects of language must rely more heavily on internal cognitive factors. Non-humans do not learn human languages, even with extensive exposure, and even within the species, adult language learners rarely achieve the fluency of children or do so in a different way (Johnson & Newport, 1989; Newport, 1990). While it is clear that both the input and internal mechanisms play a role in acquisition, the nature and contribution of each to various aspects of language learning is not fully understood.

In this paper, we investigate a part of this question, focusing on the potential contribution of the linguistic input in the acquisition of the grammatical categories Noun and Verb. The first section of this paper outlines the rationale for focusing on grammatical categories, reviews recent proposals for grammatical category learning, and introduces a distributional approach to categorization. The main body of the paper presents a series of analyses of linguistic input to young children, designed to determine whether this input contains adequate information for supporting distributional analyses of various types. We first overview our methods for performing distributional analyses, which include both computational algorithms and a set of evaluation metrics to assess how well those algorithms perform. We then present three different analyses using these methods, which vary in the aspects of input to which they have access, and we ask how well these analyses achieve approximations to the actual grammatical categories Noun and Verb. Although, we base our methods on those reported in Mintz, Newport, and Bever (1995) and Mintz (1996), a number of similarities exist between our methods and those developed by other investigators (Cartwright & Brent, 1997; Finch & Chater, 1992, 1994; Redington, Chater, & Finch, 1998). Where appropriate, we discuss the similarities and differences of these approaches. Finally, we close with a discussion of the implications of our results for theories of grammatical category acquisition, and for larger issues in language acquisition and linguistic representation.

1.1. Grammatical categories and their acquisition

The acquisition of grammatical categories—e.g., Noun, Verb—is an important aspect of acquisition because these categories are the fundamental and universal primitives from which grammars are constructed. Every language draws from a common set of lexical categories, and every language utilizes a core subset including Noun and Verb.[1] Grammatical category acquisition, thus, provides a significant test case for the evaluation of any acquisition theory, and in particular for beginning an inquiry into the importance of distributional analysis in language learning.

Two classes of theories have suggested that distributional analyses might play only a secondary role in the acquisition of grammatical categories. These theories differ with respect to the primary source of grammatical knowledge. One theory posits that *semantic* categories form the bases for grammatical categories (Bates & MacWhinney, 1979, 1982; Bowerman, 1973;

Macnamara, 1972; Schlesinger, 1974). In such a theory, the learner must observe the world to discover a word's referent and thereby its category. If it is a concrete object, for example, then the word is a noun; if it is an action or event, then the word is a verb. Different semantically based theories vary in the details of how these mappings are carried out and what the actual nature of the semantic knowledge is, but they all rely on an underlying correspondence between semantic and grammatical categories. The difficulty for such theories is that these correspondences do not always hold—for example, for nouns such as *wiggle*, *noise*, and *love*, and verbs such as *to think* and *to know* (Maratsos & Chalkley, 1980). Some investigators (e.g., Bates & MacWhinney, 1979, 1982) have proposed that the learner can generalize to non-prototypical nouns or verbs based on overlapping semantic features that they share with prototypical ones. However, it has not been demonstrated that these overlapping properties are observable in the world or that they are structured to yield the correct generalizations. Another option would be to base initial categorization on concrete referents, and then to use distributional similarities of these and their abstract counterparts to do further categorization. On this option, distributional information would play only a secondary role, and only for the more abstract items.[2]

Another class of theories (*nativist* theories) posits that the set of possible grammatical categories is innately specified (Chomsky, 1965; McNeill, 1966; Pinker, 1984). Here, the learner must still learn which words map onto which categories, but the process is highly constrained by innate knowledge of the possible syntactic structures, and thereby of possible distributional configurations of grammatical categories in utterances. Nativist theories rely on distributional analyses at some point in the acquisition process; however, the analyses are highly assisted by innate knowledge of linguistic categories and configurations. *Semantic bootstrapping* (Pinker, 1984, 1987) is a nativist theory that makes initial use of semantic–syntactic correspondences (and thus is also semantically based), which are then augmented by distributional information.

In contrast to the two types of theories presented above, some theories have proposed that *distributional information* might play a central role in categorizing words. Maratsos and Chalkley (1980) proposed that grammatical categories could be learned through a distributional analysis of the speech input. This kind of discovery procedure originated with Bloomfield (1933) and other structural linguists (Harris, 1951) who were attempting to describe how a linguist could analyze an unknown language. Grammatical categories were defined by similarities in word patterning. For instance, in sentence (1), both *dog* and *moon* are preceded by *the*, and are preceded by the same words across many sentences. This similarity would lead them to be classified together. Other words with the same pattern would be classified in the same category, and the resulting category would be nouns. Maratsos and Chalkley (1980) proposed that children might follow similar procedures in learning their native language.

(1) The dog is barking at the moon.

A number of problems have been pointed out for distributional learning theories. For example, Pinker (1987) argues that, given sentences in (2a–b), a distributional learner would incorrectly categorize *fish* and *rabbits* together and, hearing (2c), would incorrectly assume that (2d) is also permissible. Pinker argues that these erroneous generalizations would be common.

(2) a. John ate fish.
b. John ate rabbits.
c. John can fish.
d. ∗John can rabbits.

Another difficulty is that important distributional regularities are often not local, as in (1), but occur over a variable distance, as in (3) (Chomsky, 1965; Pinker, 1987).

(3) The big fluffy brown and not so thin dog is barking at the moon.

Here, the crucial co-occurrence of *the* and *dog* spans many words. The problem is how the learner could know *which* co-occurrences are important and which should be ignored. Distributional analyses that consider all the possible relations among words in a corpus of sentences would be computationally unmanageable at best, and impossible at worst.

Because of arguments like these, distributional approaches to learning grammatical categories or other aspects of syntax were abandoned. However, in recent years various researchers have again become interested in studying the distributional information available in speech, particularly in speech to infants, and have begun to show that linguistic input contains more useful information for learning grammar than was anticipated by the previous discussion (Cartwright & Brent, 1997; Mintz, 1996; Mintz, Newport, & Bever, 1995; Redington et al., 1998; Saffran, Aslin, & Newport, 1996). In the present paper, we present a series of three analyses that build on these initial findings, and that are designed to determine just how useful distributional information might be for constructing the grammatical categories Noun and Verb from linguistic input. The aim of these analyses is not to model the actual procedures a child might use, but rather to examine *the information available in children's input*. For this reason, we focus exclusively on distribution, and do not consider conceptual or semantic information at all.

In our first analysis, we use the similarity among the immediate lexical contexts of words to classify them into groups: Target words that are immediately preceded and followed by the same set of words are judged to belong to the same class. We also assess the effects on categorization of successively larger distributional contexts. These are just the kind of simple co-occurrences that should, according to arguments reviewed above, fail to represent the full structure of the grammar. Nonetheless, as we will show, nouns and verbs can be identified to a surprising degree by this type of information. Related results have previously been reported by Mintz et al. (1995), Redington et al. (1998) and Cartwright and Brent (1997), using different methods applied to child-directed speech corpora. This result also accords with Brill (1991), Finch and Chater (1992, 1994), and Schutze (1993), who used adult-directed corpora as well as different methods of analysis. This first set of results is thus a stable and widely reported finding, despite its unexpected nature. We reproduce them here with our specific corpora and procedures, to confirm the generality of the findings and to provide a performance baseline for subsequent analyses. In the central findings of the present paper (Analyses 2 and 3), we manipulate the processing mechanisms in various ways, to more closely approximate the information that very young children might perceive and represent from their input.

2. General method

In Analysis 1, we began with a "bare bones" procedure, which analyzed the distributional context of a word based only on the immediately preceding and following words. Then we extended the procedure to analyze larger distributional window sizes (two words to the right and left of a target word, and eight words to the right and left of a target word) to investigate how the domain of distributional information affects categorization. These analyses also provided a baseline categorization performance for the four corpora in our study, to be used for comparison with the results of more sophisticated analyses. Finally, for each corpus we calculated *chance* categorization results by running our procedure on random corpora generated from the tokens of each actual corpus. In Analyses 2 and 3, we modified the procedure, processing the input in two different ways, each designed to ascertain the importance of a particular aspect of the input for successful categorization. Specifically, we examined the effect of (1) imposing phrase structure boundaries to limit the distributional context, and (2) reducing the representations of elements in the input which young children might not fully process. The motivation behind these analyses was to establish whether successful categorization could result under more realistic assumptions about how young children might represent their input.

2.1. Input

By age 2.5 years, children characteristically produce utterances that display some rudimentary syntax and knowledge of grammatical categories. Therefore, some initial category learning must take place before this point. Accordingly, all of the input for our distributional analyses consists of utterances directed at children less than 2.5 years old. Input corpora were selected from the CHILDES database (MacWhinney, 2000) with the criterion that each corpus contain a substantial number of utterances directed at a child under 2.5 years old. This resulted in the following sub-corpora: Peter (Bloom, 1970; sessions 1–12; 19,846 child-directed utterances), Eve (Brown, 1973; sessions 1–20; 15,456 child-directed utterances), Nina (Suppes, 1974; sessions 1–23; 6,950 child-directed utterances), and Naomi (Sachs, 1983; sessions 1–58; 14,417 child-directed utterances). The average number of child-directed utterances across these four sub-corpora is 14,167 ($SD = 5{,}356$). The procedures and results presented below are based on individual analyses of each sub-corpus (henceforth referred to as *the corpora*).[3]

The analyses presented below are based on the 200 most frequent words within each corpus.[4] For each corpus, the 200 most frequent words account for over 80% of the tokens, and less frequent words have very low frequencies. Since the goal here was to investigate the information available in a word's various distributional contexts, low frequency words are not useful, as they have very few contexts. Table 1 shows the number of tokens of the 200 most frequent types for each corpus, the range of frequencies for the 200 most frequent words, and the number of noun and verb types in the set of target words for each corpus.

2.2. Algorithm for distributional analysis

Our categorization procedure groups words based on the similarity of their immediate lexical contexts. First, the procedure constructs a list of all the different words that appear in the corpus.

Table 1
Number of words, frequency range, and number of noun and verb types for the 200 most frequent words in each corpus

Corpus	Tokens	Frequency range	Types	
			Nouns	Verbs
Peter	74,632	72–5405	27	40
Eve	49,933	51–3374	29	39
Nina	58,221	59–3765	40	47
Naomi	22,164	23–1468	37	35

Then, for each word, it records what words come immediately before and after it throughout the corpus, and how many times. Fig. 1 presents a schematic of these data, in which $w3$ represents a target word. The lists labeled $W-1$ and $W+1$ indicate how many times each word in the corpus comes before and after $w3$, respectively.

Only the contexts of the 200 most frequent words are recorded; in addition, only the 200 most frequent words are considered in calculating a target word context. For example, if an infrequent word appears before $w3$, it will not be entered in the $W-1$ list; for that instance of $w3$ in the corpus, nothing will be entered for the preceding context. Thus, in computing the context for the instance of *likes* in *John likes port*, where *port* is an infrequent word and *John* is one of the 200 most frequent, the procedure would increase by 1 the frequency of *John* in the $W-1$ list of *likes*, but no change would be made in the $W+1$ list of *likes*.

Finally, a context is only tallied within an utterance.[5] For example, if an instance of a target word is the first word of the utterance, no entry would be made in the $W-1$ list for that instance.

Taken together, the $W-1$ and the $W+1$ lists for a given target word represent the immediate distributional contexts—both preceding and following—of that word. Since the context lists are

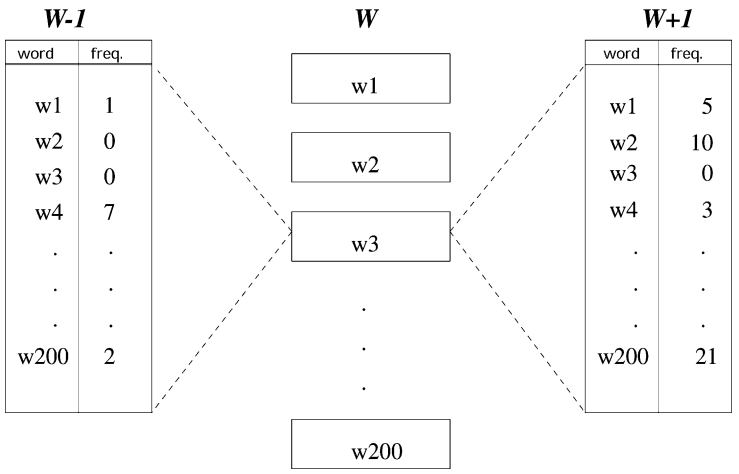

Fig. 1. Representation of distributional contexts.

ordered equivalently for all target words (e.g., following Fig. 1, the first row in $W-1$ represents the frequency of word $w1$ preceding the target word, for all target words), the contexts of two target words can be compared by comparing the concatenated $W-1$ and $W+1$ lists for these words. This can be done formally by treating the concatenation of $W-1$ with $W+1$ as a vector with twice the dimensionality of the component vectors. The context similarity of two words can then be determined by computing the angle θ between the context vectors in 400-dimensional space. Eq. (1) gives the formula for computing θ.

$$\theta = \cos^{-1}\left(\frac{\text{context\_vector}_i \cdot \text{context\_vector}_j}{|\text{context\_vector}_i||\text{context\_vector}_j|}\right) \quad (1)$$

The smaller θ is, the more similar are the two distributional contexts of the words.[6] When each target word is paired with all the other target words and their distributional contexts are compared, the result is a similarity rating for every pair of words.

After the distributional context similarity is calculated for every word pair, this information is submitted to a hierarchical cluster analysis (HCA). The HCA is a best-fit hierarchical representation of the similarity space of all the word pairs. The similarity of two words is represented by the height in the tree of their lowest common branch: the lower down they are connected, the more similar they are. Thus, similar words form clusters based on their distributional similarity. To the extent that distributional information is sufficient for inducing grammatical categories, words of the same category (such as Noun) should be clustered together by the HCA. To evaluate the success of the procedure, one must label each word with its actual linguistic category and examine how well the analysis clustered words of the same linguistic category together. Henceforth, we refer to linguistic categories as *categories* and to the categories that the analysis generates as *groups* or *clusters*.

For the analysis in which the distributional contexts extend to two words on either side of a target word (henceforth the *2-word analysis*), the procedure was modified to construct a $W-2$ and a $W+2$ context list, analogous to $W-1$ and $W+1$ in Fig. 1. For each target word, this analysis tallies how many times each word in the corpus appears in the second position to the left and in the second position to the right, as well as in the first position to the left and the first position to the right. The vector representation for each target word's context thus has twice the dimensionality as those in the 1-word analysis. Words are then paired and their context vectors compared in the same way as in the 1-word case.

For the version in which the distributional contexts extend to eight words on either side of a target word (henceforth the *8-word analysis*), the program was modified to keep track of words in all positions to the left and right of a target word, up to and including the eighth position. The resulting vector for each target word thus has eight times the dimensionality as its counterpart in the 1-word case. Words were then paired and their context vectors compared as described above.

2.3. Evaluation metrics

There are two ways in which the success of an analysis was evaluated, one qualitative and one quantitative. The qualitative method was simply to observe how well the analysis grouped words of the same category together. Since the HCA yields a hierarchy of clusters, one must

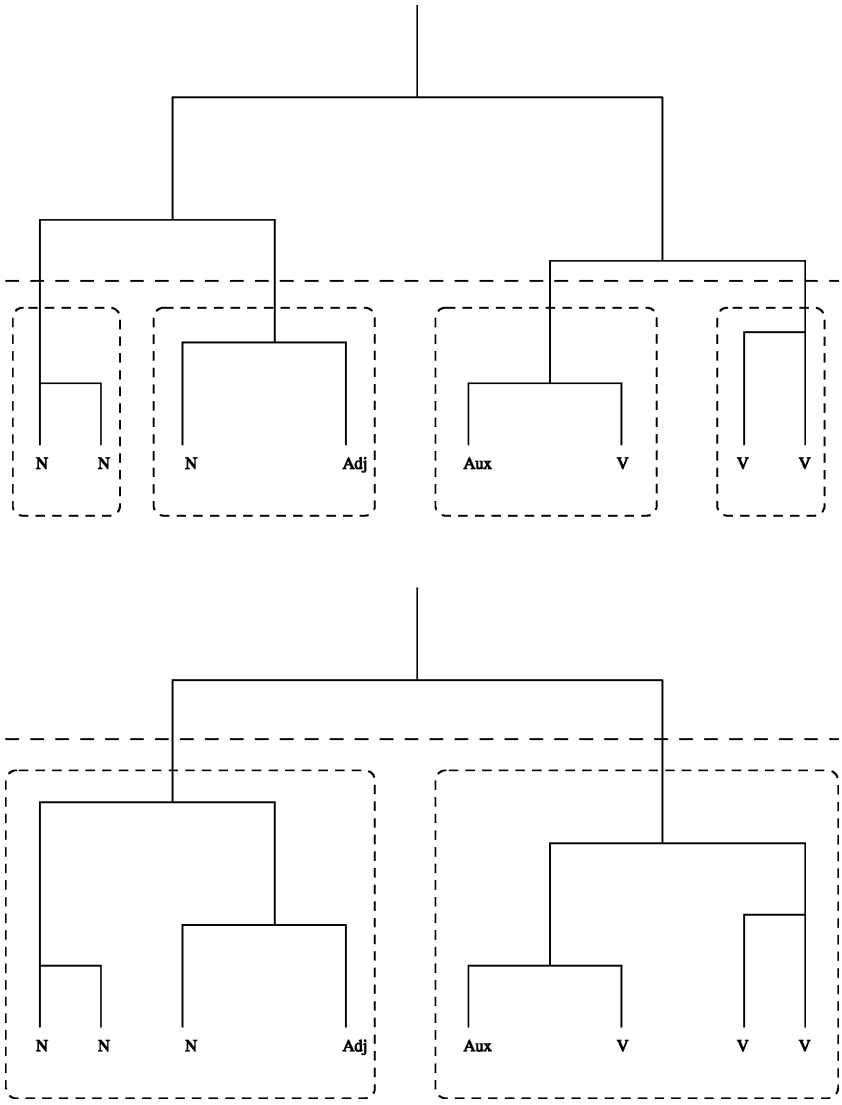

Fig. 2. Low and high thresholds for category boundaries.

pick a cutoff level in the hierarchy to obtain a specific set of clusters to evaluate. The actual cluster structure depends on the attachment height (cutoff level) one chooses. Fig. 2 shows two different clustering results based on two different attachment heights for a segment of a hypothetical HCA, along with category labels. One can see that the choice of the attachment height can influence how one interprets the success of the analysis. We return to the method used for determining cluster attachment height after we have described the method for a quantitative evaluation of the analysis.

The measure created for the quantitative evaluation of analysis is called *Purity*. Purity is calculated for each linguistic category of interest and ranges from 0 to 1. Intuitively, the Purity

of a given category X represents, for all clusters that contain a member of X, the degree to which that cluster contains purely X's; in other words, the degree to which that cluster is a pure cluster of X's. The formula for computing Purity is given in Eq. (2), where C is the number of clusters for a given attachment height, x_i is the total number of Xs in the ith cluster, w_i is the total number of words in the ith cluster, and X is as above.

$$\text{Purity} = \frac{\sum_{i=1}^{C}(x_i^2/w_i)}{X} \tag{2}$$

To calculate Purity for nouns and verbs, we needed to tag the nouns and verbs for each corpus. We made an initial classification based on our knowledge of English; for potentially homophonous words from different categories (e.g., *ride*) we consulted each corpus to see how these words were used. We categorized a potentially ambiguous word as a noun or verb if it was used as a member of that category on at least 95% of its occurrences. Only two words in our initial classifications failed this confidence test: *ride* in the Peter corpus, and *back*; these words were given neither noun nor verb classifications when calculating Purity.

As with the qualitative evaluation, the Purity score for a given category depends on the choice of the cluster attachment height. Purity increases as attachment height decreases (that is, as more numerous but smaller clusters are chosen). In the degenerate case, when each word is its own cluster, the Purity of every category is 1, because each cluster is trivially pure. Thus, having large numbers of categories (numbers close to the number of words to be categorized) yields trivially pure categories, a characteristic that might seem undesirable in a classification metric. However, as we discuss below, our evaluations are based on forming a relatively small number of substantially sized clusters. Purity is a useful measure of successful categorization under these conditions. In such circumstances, because for the most part there are many words in a cluster, Purity carries important information about how well these clusters successfully group together words of the same linguistic category. We also established that the Purity scores we obtain are not mere artifacts of the cluster attachment heights we chose (as it is in the degenerate case discussed above) by comparing the obtained Purity scores to chance Purity scores calculated from random, unstructured corpora, as we discuss below.

To determine the attachment heights on which to base our analyses, we performed pilot analyses using different ranges of attachment heights and examined the results of manipulating this variable. The pilot analyses revealed that attachment heights that yield anywhere from 20 to 30 total clusters yield on the order of 9 or 10 substantial clusters of words (clusters with at least five members); the remaining "non-substantial" clusters were generally single elements. This level of clustering assured that the number of resulting groups was constrained enough so that Purity was a meaningful metric. In addition, for both noun and verbs, Purity scores for attachment heights that yielded fewer clusters increased sharply as the number of categories increased, until around 20–30 clusters were reached (depending on the particular corpus), at which point adding clusters resulted only in a gradual increase in Purity. Thus, this range of attachment heights optimizes the conflicting goals of having a small number of clusters and having linguistically pure ones.

In separate work, Mintz (2000, 2002) developed a metric of optimal classification called Unique Entropy (UE). Based on information theoretic concepts, UE specifies the attachment height in an HCA that is potentially the most informative (essentially, the level for which

guessing the category of an item would be the hardest). Although UE is based entirely on structural properties of HCAs (not linguistic content), the attachment height range we arrived at in our pilot studies contained optimal UE clustering levels, providing independent support for our attachment height selection. Therefore, in each of the analyses presented below, the reported Purity score is the average of the Purity scores obtained from attachment heights that yielded from 20 to 30 clusters (yielding approximately 9–10 substantial clusters). The same attachment height selection criterion was applied to all corpora in all analyses, including the analyses of chance Purity (see below).[7]

2.3.1. Comparison with other measures

In related work (Cartwright & Brent, 1997; Redington et al., 1998), various different methods for evaluating categorization success were used. Cartwright and Brent (1997) used the signal detection measures *Accuracy* and *Completeness*, and Redington et al. (1998) used the information theoretic measure *Informativeness*. Below we compare Purity to these metrics at a conceptual level, and empirically for the analyses to be reported here.

2.3.1.1. Accuracy. Purity is similar to Accuracy at a conceptual level. Accuracy represents the proportion of *hits* (the number of pairs of words from the same linguistic category that are clustered together), out of the combined total of hits and *false alarms* (word pairs that are not classified together correctly). Thus, both Purity and Accuracy evaluate how well the distributionally defined clusters correspond to linguistic categories: having relatively pure clusters necessarily entails having relatively high accuracy and vice versa. However, in the standard calculation of Accuracy, a *single* Accuracy score is computed for a given clustering of words (cf. Cartwright & Brent, 1997). Since we were interested in tracking the potentially different effects of the manipulations to our analyses on noun and verb categorization, this measure was not well suited to our purposes. To make Accuracy more comparable to Purity, we computed individual Accuracy scores for nouns and verbs and averaged them across corpora in the same way as Purity scores. The correlation between these Accuracy scores and our Purity scores was significant for the full set of the analyses in this paper for nouns ($r = .88$, $p < .0001$) and for verbs ($r = .79$, $p < .0001$).

When Accuracy is modified in this way, its general characteristics are similar to Purity, but there is an important difference: For distinct clustering outcomes—e.g., from different corpora—with high Accuracy (and Purity) values, small differences between the cluster structures result in greater differences in Accuracy than in Purity (whereas at lower values, minor structure changes result in similar changes for both Accuracy and Purity). This means that for a set of slightly different analysis outcomes, the standard deviations of a category's Accuracy scores will be higher than the standard deviation of the category's Purity scores (although both measures cover approximately the same range of values). This could be important when making statistical comparisons across analyses, pooling across small numbers of corpora, as we do here. Accuracy's greater proportional variability at higher versus lower values could mask real effects that Purity might reveal.

2.3.1.2. Informativeness. Like Accuracy, Informativeness is computed for the entire category structure of a given hierarchy slice and thus was not well suited to our purposes. However,

to compared Informativeness, we computed Informativeness scores for each of our analyses. Purity significantly correlated with Informativeness for nouns ($r = .54$, $p < .00001$) and for verbs ($r = .46$, $p < .001$).[8]

Thus, for our specific goals, Purity was more suitable than Informativeness because we could independently evaluate noun and verb performance. Purity was more suitable than Accuracy because of the more sensitive comparisons that are possible across different analyses.

2.4. Chance performance

For each analysis, chance performance was determined by analyzing 10 randomized pseudo-corpora associated with each corpus. Pseudo-corpora were created from each real source corpus by stringing together tokens from the source corpus at random to form "utterances." Pseudo-corpora were created to have the same number of utterances as the associated source corpus, the same average utterance length, and the same token frequency distribution.

3. Analyses

3.1. Analysis 1: baseline categorization from adjacent word contexts

The purpose of Analysis 1 was to determine the baseline categorization performance for a distributional analysis of very local contexts (1-word windows) and for more extensive distributional environments (2- and 8-word windows) for each of the four corpora. Since the distributional contexts for this analysis are selected based on locality, not on more linguistically and structurally informative characteristics, these contexts might often exclude linguistically important distributional information and include irrelevant information. For example, two words could be immediately adjacent merely as the accidental result of two phrases being next to each other. These are the types of accidental features Chomsky (1955, 1965) and Pinker (1979, 1984, 1987) point to as potential problems for distributional analyses. While it might seem that such problems would be resolved by using wider contexts, this is not necessarily the case: while additional distributional contexts might be informative, they would also provide increased noise. Such problems would be especially serious when using relatively small corpora such as those we are using.

Nevertheless, in previous work, we and others (Mintz, 1996; Mintz, Newport, & Bever, 1995; Redington et al., 1998) have found that a distributional analysis can do surprisingly well, using both very local and wider contexts. In the present analysis, we extend these results to additional corpora for comparison to our subsequent analyses.

3.1.1. Procedure

The procedure for this analysis is outlined in Section 2. Three versions of the analysis were performed, using 1-, 2-, and 8-word windows. Table 2 provides descriptive statistics of utterance lengths for the analyzed strings from each corpus. Since the majority of utterances are not more than eight words long, the 8-word analysis includes the entire utterance as the distributional context, in most cases.

Table 2
Utterance length statistics for Analysis 1 and phrase restricted analyses (Analysis 2)

Corpus	Analysis	Mean	*SD*	Length of 95% of utterances
Peter	1	4.7	3.70	≤11
	2; Approx. phrase	1.5	0.74	≤3
	2; Percept. phrase	1.6	0.80	≤3
Eve	1	4.1	2.80	≤9
	2; Approx. phrase	1.6	0.76	≤3
	2; Percept. phrase	1.6	0.81	≤3
Nina	1	5.1	2.69	≤10
	2; Approx. phrase	1.6	0.73	≤3
	2; Percept. phrase	1.6	0.78	≤3
Naomi	1	4.2	2.86	≤9
	2; Approx. phrase	1.6	0.80	≤3
	2; Percept. phrase	1.7	0.85	≤3

3.1.2. Results and discussion

Table 3 shows the Purity values for Noun and Verb from the analysis of each of the four corpora, as well as the average over all the corpora. These results are shown for each of the three context window sizes. Each data point represents the average over the clusterings obtained from an 11-slice region of attachment heights, as described in Section 2. Shown in parentheses is the average chance Purity value for the ten associated pseudo-corpora.

For the 1-word analysis, the Purity values for nouns were quite high for all corpora. The Purity values for the verbs, while not as good as for the nouns, were also significantly and substantially better than chance (for nouns $t(3) = 24.90$, $p < .001$; for verbs, $t(3) = 13.66$, $p < .001$). Tables 4 and 5 show the actual words that were grouped together for two of the four corpora, for illustration. Results from the Peter corpus were typical, and results from the Nina corpus were the best overall.

Table 3
Purity values for different size context windows

	N			*V*		
	1	2	8	1	2	8
Peter	.76 (.34)	.78 (.35)	.78 (.35)	.64 (.28)	.56 (.28)	.69 (.28)
Eve	.75 (.32)	.75 (.34)	.77 (.30)	.66 (.32)	.65 (.30)	.71 (.32)
Nina	.76 (.32)	.74 (.31)	.78 (.32)	.74 (.37)	.74 (.37)	.80 (.38)
Naomi	.78 (.34)	.76 (.35)	.76 (.36)	.61 (.29)	.68 (.26)	.62 (.26)
Mean	.76	.76	.77	.66	.66	.71

Values in parentheses are average Purity scores for 10 associated pseudo-random corpora.

Table 4
Nina—groups based on 1-word context

animals baby ball barn bed birthday blanket book box boy car chair dog doggy doll dolly elephant eyes feet girl hair hand hat head horse house kitty lady man monkey mouse puzzle rabbit table train zoo *big blue funny good little mommy nice other red yellow*
build come doing draw eat fall feed find fit get give go going got have know let like make making open play put putting read remember say see show sit sleep take think want went *be can't don't just*
and are but can did didn't do gave if no now oh ok shall so thank when will would yeah yes
all at eating for from has here's holding in inside it's of off on or over that's there's wearing where's with
he her here his it Linda my Nina Nina's she that the there this your
color does doesn't else happened is isn't was were
back goes sleeping up what's who's
one out part picture some top
a an called not right too
I'm he's she's they're you're
him something them
how what where who why
I I'll let's we
these they those you
me to

Singletons
another down look many

Table 5
Peter—groups based on 1-word context

baby bag ball blocks book box boy car egg floor house paper pen piece see-saw slide tape top toys train truck way wheel wheels *big daddy little mommy one other*
bring close come doing eat find fix get give go going got have know leave like looking need open play push put say see show sit take think throw turn want write *be better can can't didn't does don't gonna isn't just were*
and at for here's in is it's make of on ride that's there's under what's where's who's with
are but did do help how if maybe mmhm mmm no now oh ok oops or so thank then uhhuh well what what're when where who would yeah yes
'em it something them these those
a all good nice not right too very
another he she that this
her jenny lois me patsy
I they we why you
here my the there your
I'm they're we're you're
about else shall should
I'll let's to we'll
any some two
finished yet

Singletons
awoh hm let look more recorder room time

A noticeable feature of Tables 4 and 5 is that in each case there were two groups which contained substantially more words than any of the others. One of these groups contained the majority of the nouns (24 of 27 for Peter, 36 of 40 for Nina), and the other contained the majority of the verbs (32 of 40 for Peter, 35 of 47 for Nina). We will refer to these as noun and verb groups, respectively. The relatively high Noun and Verb Purity scores reflect the fact that the groups containing nouns and verbs were not "contaminated" to a high degree by words from other categories. (Words in the Noun group which were not considered nouns for the purposes of calculating Purity, and words in the Verb group which were not considered verbs for the purpose of calculating Purity, are shown in italics.[9])

The 1-word analysis demonstrates that, even with an extremely limited distributional context of only one word to either side of a target word, an analysis based on distributional similarity has a strong tendency to cluster nouns with other nouns, and verbs with other verbs. Thus, as shown here and elsewhere (Mintz, 1996; Mintz, Newport, & Bever, 1995; Redington et al., 1998), the accidental exclusion of linguistically informative co-occurrences and accidental inclusion of irrelevant co-occurrences resulting from defining context by immediate adjacency does not severely distort the results. This must be due in part to the fact that our analysis weights *recurring* contexts, and groups together words when they share a *range* of contexts, not just one. In summary, local contexts can result in surprisingly good categorization.

Consider now the results for the 2- and 8-word analyses. As in the 1-word analysis, the Noun Purities for both the 2- and 8-word analyses were quite good. All Noun Purities were significantly and substantially better than chance (for 2-word windows, $t(3) = 72.74$, $p < .001$; for 8-word windows, $t(3) = 26.48$, $p < .001$). There was no significant difference in Noun Purity between the 1- and 2-word ($t(3) = 0.43$, $p = .70$), nor between the 1- and 8-word analyses ($t(3) = 1.06$, $p = .37$). Thus, there was no change in overall informativeness of nouns' distribution across this wide range of context windows. Any new information allowed into these expanded analyses was apparently counteracted by erroneous information about nouns' distributional privileges. However, even the 1-word purities for nouns was high, so perhaps there was little room for improvement.

Verb Purities were also significantly and substantially better than chance (for 2-word windows, $t(3) = 12.24$, $p < .005$; for 8-word windows, $t(3) = 29.86$, $p < .001$). Verb Purities were not significantly different in the 2-word analysis compared to the 1-word analysis ($t(3) = 0.12$, $p = .91$). Thus, as with nouns, there was no change in overall information about verb distribution when going from a 1- to a 2-word analysis. However, the 8-word analysis showed a significant increase in Verb Purity ($t(3) = 3.94$, $p < .03$). Table 6 shows the groups in the 8-word analysis of the Nina corpus, which showed the strongest effect.

Thus, for nouns and verbs in these corpora, it appears as though both 1- and 2-word distributional cues are equally informative for categorization. These general results accord with findings by Redington et al. (1998).

The categorization of verbs was improved by using the larger, 8-word distributional context. This was not true for nouns, which were grouped well in all analyses. That 1- and 2-word environments yielded relatively high Purities for nouns implies a stable distributional environment in the local domain of an NP (noun phrase). The slightly worse, although still quite good, results for verbs' 1- and 2-word environments suggest that distributional contexts within the VP (verb phrase), close to the verb, are more unstable. This makes sense when one considers the

Table 6
Nina—groups based on 8-word context

baby ball barn bed birthday blanket book box boy car chair dog doggy doll dolly elephant girl hat horse house kitty lady man monkey mouse picture puzzle rabbit table train zoo *big blue funny good little nice other red yellow*
build called come doing draw eat eating fall feed find fit get give go going got have holding know like make making open play put putting read remember say see sit sleep take think want wearing went *be can't don't just*
another he her here him his it Linda me mommy my Nina Nina's she some something that the them there this to your
and are but can did didn't do gave if no now oh ok so thank when will would yeah yes
all at for from has here's in inside it's of on or that's there's where's with
animals back down goes off one out part sleeping too top up
color does else happened is was were
eyes feet hair hand head
how what where who why
these they those you
I I'll let's we
a an not right
let show
doesn't isn't
Singletons look many more over

high degree of possible variability within VPs: optional modals and auxiliaries can precede the verb, and an array of prepositions can follow. Because these elements do not always appear in VP's, the distributional contexts for different types of words within the VP can overlap greatly. For example, consider the sentences in (4).

(4) a. Peter is looking in the house.
b. Peter is in the house.

The verb *looking* in (4a) and the preposition *in* in (4b) are preceded by the same sequence of words, and thus, their distributional contexts overlap to some degree. The overlap could be quite high if strings such as (4) occur often in a corpus. Indeed, one can see from Table 4 that some verbs (i.e., *eating*, *holding*, *wearing*) are grouped with a majority of prepositions. Apparently, however, the overall distributional overlap of elements within VP's is reduced when the distributional window is expanded to include 8-word information. Table 6 shows that with 8-word information, the three verbs listed above all cluster with other verbs.[10]

In sum, Analysis 1 yielded successful categorization using various context window sizes, in accord with results from other investigators using different corpora and somewhat different methods (Cartwright & Brent, 1997; Mintz, 1996; Redington et al., 1998). This analysis establishes baseline and chance Purity scores, for a variety of window sizes. Analyses 2 and 3 were aimed at achieving a better understanding of the aspects of linguistic context that are critical for these results, and at the same time examining the effects of different assumptions about the information that young children might actually use to carry out such analyses.

3.2. Analysis 2: grammatical categories from distributional contexts within phrases

We suggested above that expanding the context window might have the desirable effect of including more pertinent distributional information in the analysis, but might also include irrelevant "noise." This is one way to think about why expanding to larger contexts was not helpful overall in Analysis 1. Perhaps if there were a way to restrict the wider analyses to contexts that are more significant linguistically, the categorization results would improve. An obvious approach is to use phrase boundaries to restrict the analyses. Various researchers have suggested for natural language, and experimentally demonstrated for artificial languages, that cues to phrase boundaries are helpful in learning syntactic structure (Gleitman & Wanner, 1982; Morgan & Newport, 1981; Morgan, Meier, & Newport, 1987). Furthermore, infants have been shown to be sensitive to prosodic cues that signal phrase boundaries (Jusczyk et al., 1992), and thus, information about where phrases begin and end might be available to them. The corpora we analyze here do not contain information about prosodic features; however, phrase boundaries can be approximated by the location of closed-class words, which tend to occur at phrase edges (Kimball, 1973). Moreover, since closed-class words are typically phonologically reduced and are acoustically identifiable (Morgan, Shi, & Allopenna, 1996), even by (Shi, Werker, & Morgan, 1999), this heuristic for locating phrase boundaries is a plausible procedure which real language learners might use (cf. Gerken, Landau, & Remez, 1989; Morgan et al., 1987; Shady, Gerken, & Jusczyk, 1995). Research has shown that adults learn artificial grammars better when their input contains high frequency "function words" (Morgan et al., 1987; Valian & Coulson, 1988; Valian & Levitt, 1996) suggesting that function words might be used to segment the input. Using a different approach, Juliano and Bever (1990) and Bever (1992) provide a model in which sequential lexical distributional cues alone can be used to arrive at a phrase structure representation. Taken together, these findings suggest that even pre-linguistic infants may be able to perform phrase segmentation using acoustic, prosodic, and distributional cues.

Analysis 2 re-examines the three distributional window conditions from Analysis 1, but this time using approximated phrase boundaries to restrict the analysis. The question of interest is whether categorization—especially verb categorization—improves when distributional contexts are restricted by phrase boundaries.

3.2.1. Procedure

The procedure used in Analysis 2 is identical to the procedure used in Analysis 1, with one important difference. In Analysis 2, when a context is recorded for a given target, the existence of a closed class word within the defined distributional environment (1-, 2-, or 8-word) serves to limit that context. Thus closed-class words are treated as elements marking the front edge of phrases. This was implemented as follows: When the context to the left of a target was recorded (the *preceding* context) no context beyond the first closed-class word encountered (from right to left) was recorded, although the closed-class word itself was included in the preceding context. Similarly, when the context to the right of the target was recorded (the *following* context), no context beyond the first closed-class word encountered (from left to right) was recorded; in this case, the closed-class word was *not* recorded as part of the following context. For example, in the 2-word environment of the target word *car* in the utterance *Peter's putting the big car*

Table 7
Closed-class words

a	about	again	all	an	and	another	any
are	aren't	around	as	at	away	be	being
but	by	can	can't	could	did	didn't	do
does	doesn't	doing	don't	down	else	'em	for
from	gonna	he	he'll	he's	her	here	here's
hers	herself	him	himself	his	how	if	in
into	is	isn't	I	I'll	I'm	it	it'll
it's	itself	let's	many	me	more	my	myself
no	not	now	of	off	on	one's	or
other	our	ours	ourselves	out	over	shall	she
she's	should	so	some	someone	something	that	that's
the	their	them	themselves	then	there	there's	these
they	they're	this	those	through	to	too	up
us	wanna	was	we	we'll	we're	we've	well
what	what're	what's	what've	when	where	who	who's
why	will	with	won't	would	yeah	yes	yet
you	you'll	you're	you're	you've	your		

on the table, the preceding context would be *the big*, and the following context would be empty, since *on* is a closed-class word. The overall effect is that closed-class words serve to limit distributional contexts. In addition, the procedure effectively keeps the closed-class word within the context of its phrase (in English).

Two versions of this analysis where run, each of which used a slightly different set of closed-class words to mark phrase boundaries. In the *Approximated* version, the closed-class words include determiners, pro-forms, auxiliaries, modals, conjunctions, *wh*-words, quantifiers, and prepositions; these are listed in Table 7. In the *Perceptual* version, we used only closed-class words that we thought could reasonably be identified by perceptual characteristics, such as duration, fundamental frequency, type frequency, number of syllables (one) (see Morgan et al., 1996, for a demonstration that, combined, these cues are informative for differentiating open- from closed-class items in infant directed speech).[11] We also included non-linguistic expressions (like *uhhh*), as it is reasonable to suppose that these might be used in the same way as closed-class words to segment speech into phrases. Table 8 lists these words. Table 2 shows the effective reduction of utterance length produced by restricting the analysis of utterances to phrases.

3.2.2. Results and discussion

Table 9 shows the Purity values for Nouns and Verbs for the 1-, 2-, and 8-word distributional windows, when limited by the Approximated and Perceptual phrase boundaries, as well as the comparable Purity values from Analyses 1 and 2 when not limited in this way (henceforth referred to as the All-Words analyses). Clearly, since most phrases in this analysis are less than four words long (see Table 2), the 8-word window is more restricted than in Analysis 1, and generally corresponds to an entire phrase. We discuss the results for each version of phrase restriction in turn.

Table 8
Perceptually identifiable closed-class words

a	all	an	and	any	are	aren't	as
at	be	but	by	can	can't	could	did
didn't	do	does	doesn't	doin'	don't	down	'em
for	from	gonna	he	he'll	he's	her	here
here's	hers	him	his	hm	hmm	how	huh
humm	I	I'll	I'm	if	in	into	is
isn't	it	it'll	it's	let's	many	me	mmm
my	no	not	of	off	on	one's	or
our	ours	out	shall	she	she's	should	so
some	that	that's	the	their	them	then	there
there's	these	they	they're	this	those	to	uh
um	up	us	wanna	was	we	we'll	we're
we've	what	what're	what's	what've	when	where	where's
who	who's	why	will	with	won't	would	yeh
yet	you	you'll	you're	you've	your		

Table 9
Purities for closed-class phrase boundaries & All-Words analyses; all window sizes

Window size	Corpus	Closed-class phrase restricted				All-Words[a]	
		Approximated		Perceptual			
		N	*V*	*N*	*V*	*N*	*V*
1	Peter	.85 (.31)	.83 (.39)	.81 (.29)	.78 (.36)	.76	.64
	Eve	.73 (.31)	.80 (.38)	.68 (.30)	.76 (.33)	.75	.66
	Nina	.91 (.39)	.77 (.42)	.85 (.36)	.75 (.41)	.76	.74
	Naomi	.85 (.41)	.88 (.36)	.77 (.38)	.83 (.34)	.78	.61
	Mean	.83	.82	.78	.76	.76	.66
2	Peter	.83 (.30)	.83 (.39)	.86 (.29)	.78 (.35)	.78	.56
	Eve	.75 (.31)	.80 (.37)	.66 (.29)	.79 (.36)	.75	.65
	Nina	.76 (.38)	.75 (.42)	.78 (.35)	.83 (.41)	.74	.74
	Naomi	.77 (.41)	.83 (.36)	.69 (.39)	.75 (.35)	.76	.68
	Mean	.78	.80	.75	.79	.76	.66
8	Peter	.85 (.31)	.85 (.39)	.78 (.28)	.82 (.35)	.78	.69
	Eve	.83 (.31)	.84 (.37)	.71 (.29)	.78 (.36)	.77	.71
	Nina	.85 (.38)	.81 (.42)	.80 (.35)	.82 (.41)	.78	.80
	Naomi	.81 (.42)	.89 (.36)	.70 (.39)	.76 (.34)	.76	.62
	Mean	.83	.85	.75	.79	.77	.71

Values in parentheses are average Purity scores for ten associated pseudo-random corpora.
[a]Data taken from Table 3.

For the Approximated-Boundary analysis, as in previous analyses, Purity was significantly better than chance for nouns (1-word analysis: $t(3) = 16.31$, $p < .001$; 2-word analysis: $t(3) = 11.20$, $p < .005$; 8-word analysis: $t(3) = 13.80$, $p < .001$) and for verbs (1-word analysis: $t(2) = 12.37$, $p < .005$; 2-word analysis: $t(3) = 13.74$, $p < .001$; 8-word analysis: $t(3) = 16.12$, $p < .001$). Nouns showed a trend towards improvement in the 1-word and 2-word conditions compared to the All-Words analysis (1-word: $t(3) = 2.06$, $p = .066$, one-tailed; 2-word: $t(3) = 1.80$, $p = .085$, one-tailed) and a significant improvement in the 8-word condition ($t(3) = 15.99$, $p < .001$, one-tailed). Thus, Noun Purity not only remains constant over a range of context scopes (Analysis 1), it also stays constant or improves when contexts are restricted by easily recognized phrase boundaries. For verbs, at all window sizes there was a significant increase in Purity scores with Approximated Boundaries (1-word: $t(3) = 3.21$, $p = .025$, one-tailed; 2-word: $t(3) = 2.63$, $p = .04$, one-tailed; and 8-word: $t(3) = 2.73$, $p = .036$, one-tailed). Table 10 shows the clusters formed with the 1-word Approximated-Boundary analysis of the Peter corpus, which showed typical effects. (This and all following cluster tables are produced using the same attachment height as in Table 5.)

Results were similar for the Perceptual-Boundary condition. All Purity scores were significantly better than chance for the nouns (1-word: $t(3) = 16.91$, $p < .001$; 2-word: $t(3) = 7.28$, $p < .01$; 8-word: $t(3) = 10.45$, $p < .005$) and for the verbs (1-word: $t(3) = 13.08$, $p < .005$; 2-word: $t(3) = 59.40$, $p < .001$; 8-word: $t(3) = 31.76$, $p < .001$). Noun Purity was unchanged compared to the All-Words analysis (1-word: $t(3) = 0.39$, ns; 2-word: $t(3) = 0.30$, ns; 8-word: $t(3) = 1.30$, ns); however, verb Purity showed significant improvement for all window sizes compared to the All-Words analysis, although the magnitude of these changes was not as great as in the Approximated-Boundary condition (1-word: $t(3) = 3.19$, $p = .025$, one-tailed; 2-word: $t(3) = 3.52$, $p = .02$, one-tailed; 8-word: $t(3) = 3.14$,

Table 10
Peter—approximated boundary, 1-word context

baby bag ball blocks book box car egg floor house paper pen see-saw slide top toys train truck wheel wheels *daddy mommy*

bring close come eat find fix get give going got have know leave let like look looking make need open play push put say see show sit take thank think throw turn want write *maybe under were*

I I'll I'm a about again all and another any are around at away be but can can't did didn't do does doesn't doing don't down else 'em for gonna he her here here's how if in is isn't it it's let's me more my no not now of off on or other out over shall she should so some something that that's the them then there there's these they they're this those to too up was we we'll we're well what what're what's when where where's who who's why with would yeah yes yet you you're your

awoh hm huh mmhm mmm oh ok oops uhhuh

jenny lois patsy pete peter right

help piece ride time

home together

one way

good nice

back goes

Singletons
better big boy finished go just little recorder room tape two very

Table 11
Peter—perceptual boundary, 1-word context

baby bag ball blocks book box car egg floor house paper pen see-saw slide top toys train truck wheel wheels *daddy mommy*
bring close come eat find fix get give got have know leave let like look looking make need open over play push put see show sit take thank think throw turn want write *maybe over under well were*
I I'll I'm a all and any are at be but can can't did didn't do does doesn't don't down 'em for gonna he her here here's hm how huh if in is isn't it it's let's me mmm my no not of off on or out shall she should so some that that's the them then there there's these they they're this those to up was we we'll we're what what're what's when where where's who who's why with would yet you you're your
finished help jenny lois now patsy pete peter piece ride something time
awoh mmhm oh ok oops two uhhuh yeah yes
again away boy home together too
better doing going say
good nice right
around back goes
about else
one way
Singletons another big go just little more other recorder room tape

$p = .026$, one-tailed). Table 11 shows the clusters formed with the Peter corpus in the 1-word Approximated-Perceptual analysis.

These results suggest that, as we predicted, restricting the distributional analyses to phrasal units can increase their informativeness. Perhaps most important, both noun and verb categorization achieved by the Approximated-Boundary analysis with only 1-word contexts is *better* than that achieved by the much wider 8-word context of Analysis 1. Similarly, categorization in the Perceptual-Boundary analysis with only 1-word contexts was at least as good as the All-Words analysis with 8-word contexts. These results suggest that, if a learner has access to some rough approximation of phrase boundaries, a very local distributional analysis could yield very good noun and verb categories, and with a more precise identification of phrase boundaries, the classifications would improve. Restricting the analysis by either method would place a far less substantial computational and representational load on the learner while achieving a similar outcome. When we consider what distributional patterns infants and very young children might realistically be able to keep track of, it seems likely that local distributional information would be more available to them than distant information (see Santelmann & Jusczyk, 1998), so this is a welcome result.

Thus, Analysis 2 demonstrates that the results of Analysis 1 can be surpassed by restricting distributional contexts to phrases. Furthermore, relatively local contexts can be *more* informative than large contexts when the local contexts do not cross phrases boundaries.

Although the results of the previous analyses are quite good, our procedure represents all occurrences of the 200 most frequent words and all of these words as contexts. This requires considerable mnemonic resources, even if only relatively local distributional environments are considered. In the next analysis, we investigate the effect of severely limiting the information the procedure incorporates, by reducing all closed-class words to the same type.

3.3. Analysis 3: grammatical categories from distributional contexts with reduced closed-class words

The procedures used in the preceding analyses keep track of the 200 most frequent words as both targets and as parts of the distributional contexts. To better approximate the information available to infants and young children, it may be more accurate to fully represent only a subset of the 200 most frequent words and their distributional contingencies. It has been suggested that, although speech produced by 2 years old lacks function words, young language learners might nonetheless represent them, perhaps in a reduced or undifferentiated form (Echols, 1993; Gerken et al., 1990; Gleitman & Wanner, 1982; Newport, Gleitman, & Gleitman 1977; Peters, 1977; Shipley, Smith, & Gleitman, 1969). More recent research indicates that even infants have representations of function words that are more detailed, allowing them to differentiate real function words from nonsense ones, and have expectations of where function words occur in sentences (Höhle and Weissenborn, 1998; Shady, 1996; Shafer, Shucard, Shucard, & Gerken, 1998). However, in tracking the distributional properties of function words, it is unclear whether infants differentiate between particular function words or treat them uniformly, at least until approximately 16 months of age (Shady, 1996). Therefore, we chose to investigate the effects on distributional analyses of the extreme case in which the closed-class words are represented but are not differentiated from each other at all. Thus, in this analysis we constrained the procedure's representation of closed-class words, by reducing them all to the same type.

Collapsing the representations of closed-class words provides a way of testing the robustness of the distributional information available in the input. The grammatical nature of closed-class words is such that they may have provided much of the distributional structure that yielded the relatively good categorization results so far. For example, nouns are preceded frequently by articles—*a* or *the*—and only sometimes by adjectives or other determiners. One might imagine that by representing articles in the same way as other closed-class words (like auxiliaries and modals), we would undermine the structural information that the previous results relied on. However, the differences between a number of environments, like those containing prepositions, would disappear: sequences such as . . . *put on* . . . and . . . *take off* . . . would reduce to . . . *put FNCT* . . . , and . . . *take FNCT* . . . , which renders the distributional environments of these two verbs more similar. *A priori*, the net result of these potentially opposing effects is unclear.

Since closed-class words are among the most frequent words for each corpus, a side effect of collapsing their representation is a reduction in the memory requirements for the procedure. The dimensionality of each word's context vector is reduced, and the number of words to classify is reduced. For these corpora, Analysis 3 uses approximately 75% less memory than the full analyses to track co-occurrence patterns.

Thus, this analysis provides an empirically motivated way to explore how a distributional analysis stands up to the degradation of potentially very important distributional cues (e.g., the difference between articles and modals), as well as to demonstrate what kind of categorization can be achieved under greater constraints on computational resources. As in Analysis 2, we perform two versions of this analysis, using the Approximated and the Perceptual specifications of closed-class words. In the Approximated version, we address these issues "in principle," as they pertain to the linguistically defined set of function words. In the Perceptual version,

we address these issues as they might pertain to human learners, using reasonable perceptual characteristics to identify closed-class words.

3.3.1. Procedure

The procedure used in Analysis 3 is similar to that of Analysis 1, with one important difference. In Analysis 3, individual closed-class words were not included in the representation of a target word's distributional context. Instead, all closed-class words were replaced by a single symbol. For example, using either an Approximated or Perceptual definition of closed-class words, *Peter's putting the big car on the table* becomes *Peter's putting FNCT big car FNCT FNCT table*, where FNCT replaces all closed-class words. Another difference from Analysis 1 is that closed-class words were not included in the set of words to be categorized, so just over 100 words (rather than 200 words) are classified for each corpus.

3.3.2. Results and discussion

Table 12 shows the Purity values for nouns and verbs for Analysis 3, for both Approximated and Perceptual closed-class words. For comparison, the original Purity values for Analysis 1 are also given. As in previous analyses, all Purity scores for each window size were better than chance for nouns (1-word Approximated: $t(3) = 9.24$, $p < .005$; 2-word Approximated:

Table 12
Purities for closed-class word reduction & All-Words analyses; all window sizes[a]

Window size	Corpus	Closed-class words reduced				All-Words[a]	
		Approximated		Perceptual			
		N	*V*	*N*	*V*	*N*	*V*
1	Peter	.79 (.50)	.83 (.50)	.65 (.46)	.76 (.46)	.76	.64
	Eve	.74 (.47)	.83 (.48)	.63 (.43)	.79 (.45)	.75	.66
	Nina	.74 (.50)	.76 (.54)	.73 (.47)	.74 (.50)	.76	.74
	Naomi	.67 (.50)	.85 (.42)	.64 (.46)	.75 (.38)	.78	.61
	Mean	.73	.82	.66	.76	.76	.66
2	Peter	.71 (.47)	.84 (.51)	.62 (.44)	.73 (.45)	.78	.56
	Eve	.78 (.48)	.77 (.48)	.65 (.45)	.81 (.45)	.75	.65
	Nina	.72 (.51)	.77 (.55)	.68 (.47)	.72 (.52)	.74	.74
	Naomi	.66 (.50)	.77 (.42)	.63 (.47)	.71 (.39)	.76	.68
	Mean	.72	.79	.65	.74	.76	.66
8	Peter	.85 (.47)	.86 (.51)	.70 (.47)	.83 (.45)	.78	.69
	Eve	.75 (.47)	.84 (.50)	.69 (.42)	.83 (.45)	.77	.71
	Nina	.72 (.50)	.78 (.55)	.72 (.46)	.79 (.51)	.78	.80
	Naomi	.72 (.50)	.83 (.42)	.68 (.47)	.82 (.40)	.76	.62
	Mean	.76	.83	.70	.82	.77	.71

Values in parentheses are average Purity scores for ten associated pseudo-random corpora.
[a]Data taken from Table 3.

$t(3) = 7.78$, $p < .005$; 8-word Approximated: $t(3) = 7.29$, $p < .01$; 1-word Perceptual: $t(3) = 21.37$, $p < .001$; 2-word Perceptual: $t(3) = 16.91$, $p < .001$; 8-word Perceptual: $t(3) = 17.61$, $p < .001$) and for verbs (1-word Approximated: $t(3) = 7.68$, $p < .005$; 2-word Approximated: $t(3) = 10.37$, $p < .005$; 8-word Approximated: $t(3) = 8.67$, $p < .005$; 1-word Perceptual: $t(3) = 11.12$, $p < .005$; 2-word Perceptual: $t(3) = 8.49$, $p < .005$; 8-word Perceptual: $t(3) = 12.22$, $p < .005$).

Comparisons to the All-Words analyses used two-tailed t-tests, since we had no prediction about the direction of the effect. When the Approximated set of closed-class words was reduced, noun Purity scores were not significantly different from the All-Words analysis, for any window size, although the values were numerically slightly lower (1-word window: $t(3) = 1.03$, $p = .38$; 2-word window: $t(3) = 1.43$, $p = .25$; 8-word window: $t(3) = .53$, $p = .63$). However, verb Purity in the 1-word analysis showed a significant *improvement* ($t(3) = 3.34$, $p < .05$). As in Analysis 2, the Purity for nouns and verbs in 1-word condition of this analysis is on a par with the 8-word version condition of the All-Words analysis. Verb Purities did not change significantly with closed-class word reduction in the 2- and 8-word analyses (2-word window: $t(3) = 2.32$, $p = .10$; 8-word window: $t(3) = 2.38$, $p < .098$); but for the 2-word analysis, Purities ranged from .77 to .85 when closed-class words were reduced, versus .56 to .74 for the All-Words analysis. In fact, all verb Purities for each corpus were higher in this analysis, compared to the All-Words analysis. In the 8-word condition Purities for closed-class reduction ranged from .78 to .88, versus .62 to .80 in the All-Words analysis, and verb Purities were higher for three of the four corpora. Table 13 shows the groups formed by the 1-word Approximated-Reduced analysis for the Peter corpus.

When the Perceptual set of closed-class words was reduced, Noun Purity decreased significantly in the 1-, 2-, and 8-word analyses (1-word window: $t(3) = 4.18$, $p = .025$; 2-word window: $t(3) = 5.14$, $p = .01$; 8-word window: $t(3) = 18.27$, $p < .001$); but nonetheless the resulting Purity scores were still quite good. As when Approximated closed-class words were reduced, there was a trend towards higher verb Purities compared to the All-Words analyses (Analysis 1), but for no window size was this improvement significant (1-word window: $t(3) = 3.10$, $p = .054$; 2-word window: $t(3) = 1.73$, $p = .18$; 8-word window: $t(3) = 2.46$,

Table 13
Peter—approximated-reduced analysis, 1-word context

baby bag ball book box car egg floor house paper pen see-saw slide train wheel *daddy*
bring close eat find fix get give goes got have know leave like looking make need open play push put show sit take think throw turn want *were*
come let look maybe thank
back blocks finished go going help lois mommy one patsy piece ride right say see top toys way wheels write
awoh hm mmhm mmm ok oops uhhuh
jenny pete peter
big just little
together truck
Singletons
better boy good home huh nice oh recorder room tape time two under very

Table 14
Peter—perceptual-reduced analysis, 1-word context

baby bag ball book box car egg floor house paper pen see-saw slide train way wheel *daddy doing mommy more one right too*
bring close eat find fix get give goes got have know leave like looking make need open play push put see show sit take think throw turn want *about were*
back blocks else finished go going help lois patsy ride say top toys wheels write
awoh mmhm oh ok oops uhhuh yeah yes
come let look over under
jenny now pete peter
piece something time
just little other
again together truck
maybe thank well
good nice
Singletons
another around away better big boy home recorder room tape two very

$p = .09$). However, Purity scores for each corpus were higher in the 1-word condition in this analysis compared to the All-Words analysis (Analysis 1), and three out of four were higher in the 2-, and 8-word conditions, when compared to the All-Words analyses. Table 14 shows the groups formed by the 1-word Perceptual-Reduced analysis for the Peter corpus.

The important result from this analysis is that, even with a reduced representation of closed-class words as parts of the distributional context, there was little qualitative degradation in categorization (see Tables 13 and 14). In fact, verb classification either stayed the same or improved. Quantitatively, noun classification declined slightly but remained high.[12]

The assumptions motivating Analyses 2 and 3 might seem to be at odds. In Analysis 2, closed-class words were fully differentiated and used in a special way, whereas in Analysis 3, they were collapsed into one word type. However, these analyses were intended to explore different aspects of infants' and young children's input representations that are supported by the empirical literature. The manipulation in Analysis 3 was extreme, in that it probably underestimated the degree to which learners above 16 months of age are able to distinguish individual closed-class words and their distinct distributional privileges. However, the analysis demonstrated the robustness of the distributional information even without distinct representations among closed-class words, showing that this information could potentially be used by learners early on. Closed-class words were used in Analysis 2 as a practical way of identifying phrase boundaries given the lack of acoustic information in the corpora, although both acoustic cues to phase boundries, and the placement of function words have been shown to be available to young learners. In practice, some mixture of these representations and processes are probably at play. Some closed-class words might anchor distributional boundaries; others might be underrepresented place-holders. Here, we have shown that either extreme yields a fairly accurate rendering of noun and verb categorization from infant directed speech.

4. General discussion

The goal of this paper was to determine what information might be available in speech to young children from which they could learn the grammatical category structure of their native language. To that end, we conducted a series of distributional analyses of speech directed at children under 2.5. In accord with related investigations (Cartwright & Brent, 1997; Mintz, 1996; Mintz et al., 1995; Redington et al., 1998), we found that this type of input can provide enough distributional information to induce, at least roughly, the major grammatical categories Noun and Verb. We further found that 1-, 2- and 8-word contexts were equally informative for noun categorization, whereas verbs benefited from 8-word contexts. Going beyond our initial results, we found that improved verb categorization occurred when operationally defined phrase boundaries restricted the domain of analysis. Even analyses with 1-word windows, restricted in this way, achieved noun and verb classification equal to full 8-word window analyses. Perhaps the most surprising result is that, even when closed-class words were not differentiated, the distributional analysis did not catastrophically degrade, and in the case of verbs actually improved. This is a promising result, given the limited resources of very young learners.

Let us now reconsider prior arguments against distributional approaches to grammatical category learning. Consider first the type of problem raised by Pinker (1987) in (2), repeated here. The distributional facts in (2a–c), Pinker claims, would lead a distributional learner to treat *rabbits* as a verb, and thus (2d) would erroneously be judged grammatical.

(2) a. John ate fish.
b. John ate rabbits.
c. John can fish.
d. ∗John can rabbits.

However, a statistically based distributional analysis would not be subject to this problem. If *fish* were the only word in the corpus to share any distribution with *rabbits*, then indeed *rabbits* would be assigned the same categorization as *fish*, i.e., it would be incorrectly classified as both a noun and a verb. But in any sizable corpus of actual speech directed at very young children, *rabbits* will likely share many more distributional characteristics with words used primarily as nouns. This would statistically override any marginal effects produced by sequences like (2a–c). To take a real example from the speech directed at Peter, the word *ride* is used approximately 70% of the time as a verb and the remainder of the time as a noun. Nonetheless, nouns that appear in similar distributional frames as the noun *ride* (for example, *toys*) were not miscategorized as verbs.[13] Thus, at least for the child directed corpora we analyzed, Pinker's argument does not apply. Apparently, the frequencies of sequences that could lead to erroneous generalizations are low enough, compared to those that lead to correct ones, as to have little effect on categorization outcomes.

Consider now the potential problem raised by utterances like the one in (3), repeated here:

(3) The big fluffy brown and not so thin dog is barking at the moon.

In simple noun phrases, the noun will usually be directly preceded by an article, as in *the moon* in (3). But in more complex noun phrases, the article and the target noun can be separated by a variable number of words. We suggested that a distributional analysis could have difficulty

with examples like (3), because *the* precedes a word that is not a noun, and the noun *dog* is not preceded by *the*. Shouldn't this cause *big* to be categorized with nouns? Shouldn't *dog* fail to be categorized with nouns? In fact, we observed some problems of the first sort, but very few of the second. For example, Tables 4 and 6 show that some adjectives were grouped with nouns, and some non-verbs internal to the verb phrase were grouped with verbs. However, this is not devastating for distributional analyses, which could still carry out a significant amount of work in carving out initial categories. Below we discuss how these initial representations could be refined. A likely explanation of the fact that the second type of problem—where *dog*, in (3), is not grouped with other nouns—did not arise is that there are distributional frames other than [ARTICLE NOUN] that could lead a noun to be grouped with other nouns. For example, in (3), *dog* is preceded by an adjective and followed by a verb, as are other nouns. These overlapping distributional cues appear to be adequate to group nouns and verbs correctly.

4.1. Other sources of information

Although the categorization results that were achieved were good approximations to actual linguistic categories, they were not perfect. Although nouns and verbs were for the most part appropriately grouped, some groups that were formed had no obvious linguistic relevance. In addition, although noun and verb groups were quite pure, syntactic relatives of nouns (e.g., adjectives) were sometimes confused with nouns, and syntactic relatives of verbs (e.g., modals, auxiliaries and adverbs) were sometimes confused with verbs. This problem presumably arises because the misclassified words often appear in similar distributional contexts to their syntactic heads (i.e., nouns and verbs). Although larger corpora (of the size that an actual learner might receive) might provide enough additional distributional information to eradicate these misclassifications, other sources of information could also be called upon to refine these analyses. Below we briefly discuss possible sources of additional information and how they might be used to refine the noun and verb categories created by the analyses presented here.

4.1.1. Semantic/conceptual information

In the introductory paragraphs, we outlined some reasons why entirely semantically driven proposals of grammatical category acquisition are problematic. However, taken together with distributional information, referential properties of words could be helpful in correcting erroneous classifications. Such a system may seem reminiscent of *semantic bootstrapping* (Pinker, 1984), with a "re-ordering" of the same sources of information—distributional and semantic. While there are indeed similarities in this respect, there are also important differences: (1) traditionally, semantic bootstrapping assumes innate knowledge of grammatical categories (i.e., that there are nouns and verbs to be found) and innate knowledge of how the grammatical categories typically map onto semantic categories; (2) the distributional analyses which are required to allow learning of semantically atypical words (e.g., abstract nouns) are assumed to be highly constrained by innate syntactic knowledge of phrase structure. While our data do not allow any conclusions about how the two putative information sources might be integrated, they do suggest that the distributional component could proceed without appeal to syntactic knowledge of phrase structure, and perhaps without innate knowledge of grammatical categories. Furthermore, our demonstration that the results of distributional analyses

come quite close to actual syntactic categories allows for the interesting possibility that core semantic–syntactic correspondences might be learned, rather than innately specified.[14] If this were so, then semantic bootstrapping would stipulate very specific linguistic knowledge for a task that could potentially be accomplished by more general operations on the linguistic input. Although the empirical work to address these issues is at a very early stage, we would suggest that an approach like the one presented in this paper can be useful for sorting out these very questions.

4.1.2. Morphology

In the analyses presented here, words are treated as atomic units. However, many words contain internal morphemes. For English, bound inflectional morphemes might be used to refine or augment a distributional analysis. Indeed, inflectional morphology was part of Maratsos and Chalkley's (1980) conception of what the important distributional units might be. Distributional regularities at this level could be used to refine the initial groupings derived from lexical sequential information, or they might contribute to categorization from the beginning. There is evidence that by 18 months infants are sensitive to the correlation of *is* and *-ing* in English progressive sentences (Santelmann & Jusczyk, 1998), so clearly by this age learners are picking up on the distributional regularities of some sub-lexical morphemes. Thus, bound morphemes could provide an additional source of distributional information about a word's grammatical category.

4.1.3. Syntax

A learner could also use rudimentary knowledge about syntactic phrases to reclassify misclassified words. For example, English nouns can occur within a noun phrase without adjectives, but the reverse is not the case; likewise, main verbs can occur in a verb phrase without modals and auxiliaries, but the reverse is not the case. This distinction between optional and obligatory categories, and the formal notion of phrase and phrasal head is core syntactic knowledge that could prevent optional modifiers from being classified with heads of phrases with which they share highly overlapping distributional contexts. Whether this knowledge must be innately specified or whether it could arise from operations on the input is an open question.

Thus, the coalition of distributional information, rudimentary syntactic knowledge, and semantic primitives could go a long way in determining the structure of grammatical categories during acquisition. We propose that distributional information might play a very early role.

5. Conclusion

We presented here a series of analyses of young childrens' linguistic input to determine what information it plausibly offers to grammatical category learning. In accord with Mintz et al. (1995), Mintz (1996), Cartwright and Brent (1997) and Redington et al. (1998), we showed that, given a learner who is predisposed to calculate distributions over words, the input contains information from which the grammatical categories of at least nouns and verbs could be constructed. Furthermore, we showed that these results are robust under a range of assumptions about the speech elements represented by infants and young children.

Robust distributional information could play a role within a number of theories. On a strongly nativist view, the present results could be interpreted as evidence that even a very limited distributional analysis could provide the basis for mapping words onto innate grammatical categories. On a less nativist view, these results could be taken as preliminary evidence that grammatical categories themselves might be constructed from distributional analyses, perhaps in concert with other information sources. However, even on the latter view, these results do not show that grammatical categories can be learned without any prior knowledge. Rather, they point to the kinds of constraints a learner would need to have in order to benefit from distributional co-occurrence information. Minimally (1) the system must be predisposed to carry out distributional analyses on repeated sequential phenomena in its environment; (2) the system must be predisposed to construct categories based on these analyses; (3) to the degree that other information sources are incorporated into the refinement or construction of these categories, the way these information sources are integrated must be specified as part of the learning mechanism; and (4) these categories must become "grammatical" (i.e., they must include structural information, not only typical distribution or reference). To that end, learners must have procedural information about what kinds of information to seek to determine the lexical category of a given distributional group.

One might conclude that we have reduced the nativist claim that lexical categories are "innate." However, what we have done is importantly different: we have specified what particular kinds of knowledge must be available to the learner to account for the extraction of a syntactically categorized lexicon. This is useful in making the nativist claim more precise and testable. For example, evidence is developing that infants have a predisposition to group speech stimuli distributionally (Gomez & Gerken, 1999; Mintz, 1996; Saffran et al., 1996).

In our view, applying techniques like those developed here to other levels of linguistic representation may lead to a better understanding of how innate predispositions combine with structural information in the input during language acquisition. The work presented here is a preliminary step, but is the kind of investigation that we think might be fruitful for addressing these issues in other aspects of grammar. While we cannot prejudge the results, such analyses may lead to further insights about how acquisition works, and may further specify the computational power required to extract grammatical information from the input and the structural universals required to interpret what is extracted.

Notes

1. Though see Jelinek (1995), who suggests that Salish may have only a contrast between functional and lexical categories, and no distinction between nouns and verbs. Nonetheless, the Noun/Verb contrast is pervasive in the world's languages.
2. A more general problem for semantic theories is that the learner must know the intended referent of the uncategorized word. The difficulty of this becomes apparent when one considers that a sentence describing the event of a boy walking could contain the word *walking* (a verb) or the words (*a*) *walk*, *action* or *motion* (all nouns). This is the same problem that theories of lexical acquisition must solve, and most researchers in this area agree that the mapping between word and referent must be mediated by representations

which are not themselves part of the referred-to world (cf. Gillette, Gleitman, Gleitman, & Lederer, 1999; Gleitman, 1990; Pinker, 1984; Quine, 1960).

3. In contrast, Redington et al. (1998) pool data from all corpora together, regardless of the age of the children. Cartwright and Brent (1997) analyze the combined speech of nine mothers.
4. Since some words are precisely tied for frequency at the lower bound of this range, it is not always possible to select *exactly* 200 words. For each of the four corpora, then, the closest possible number to 200 words is used. For the Peter corpus this is 197 words, for the Eve corpus 199, for the Nina corpus 198, and for the Naomi corpus 200.
5. Redington et al. (1998) explicitly test the effect of limiting a distributional analysis to utterance units, as opposed to allowing distributions to be tallied across utterance boundaries.
6. Because an angle is used as a measure of similarity—the smaller the angle, the more similar the contexts—the relative lengths of the context vectors are not reflected in the similarity measure. For example, if two target words appear in identical contexts, but one target word appears twice as often as the other, their context vectors will exactly overlap but one will be twice as long as the other. However, since the angle between the vectors is 0, the word contexts are rated to be identical.
7. Although, we base our statistical analyses on the averages across the range of attachment heights, the same pattern of qualitative and statistical results as those we report is found for any specific attachment height in this range. This is because there is very little variability in the resulting cluster structures in this range.
8. In order to calculate informativeness we had to supply linguistic category labels for all the 200 target words in our corpora, not just nouns and verbs. We performed these classifications in a similar manner to the way we classified nouns and verbs, as described in Section 2.
9. Note that proper names, including *mommy* and *daddy*, were not counted as nouns in these analyses. Our noun Purity measure is therefore somewhat conservative. Modals, auxiliaries, and contractions with negation were not counted as verbs.
10. Redington et al. (1998) carry out a similar set of analyses to assess the effect of the distance of context words from target words on categorization. On the whole their findings are similar to ours in that they fail to find an advantage of more distal contexts. However, because they did not analyze separately the effect of context distance on noun and verb categorization, the difference we hypothesize in immediate distributional stability between NPs and VPs was not apparent.
11. Strictly speaking, type frequency and number of syllables are not perceptual characteristics. We use the term *perceptual* as a convenient label. We also include several disyllables like *gonna*, and contractions with negation.
12. These results are in accord with earlier qualitative findings by Mintz et al. (1995), in an analysis in which they replaced functions words with a unique symbol. In a subsequent study, Redington et al. (1998) evaluated how substituting a category label for words in one linguistic category would affect the categorization of the remaining words. Although their focus was different than that of the present analysis, one part of their results involved substituting closed-class words with a category symbol and

therefore can be compared with our Analysis 3. Contrary to our results, Redington et al. (1998) found a decrement in categorization from this substitution. One difference between the two analyses is the age of the children to whom the input was addressed. Whereas we limit our corpora to speech to children under 2.5 years of age, the corpora used by Redington et al. (1998) include speech directed at older children as well. The distributional patterns of the function words might be somewhat different in the two corpora for this reason. We do not believe the difference in results is due to the different evaluation metrics we use; as mentioned above, when we calculated an *Informativeness* score following Redington et al. (1998) for the entire suite of our results, the outcomes correlated highly with both our noun and verb Purity scores.

13. The issue of how multicategory words are successfully classified is an important one. Although the procedure developed here does not make multiple classifications of a single word type, the procedure proposed by Cartwright and Brent (1997) can assign a word to more than one category based on distributional information.
14. While such learning might be fairly straightforward for nouns, some rudimentary knowledge of predicate-argument structure might be necessary for verbs (Gleitman, 1990).

Acknowledgments

This research was supported in part by a NIH National Research Service Award (MH10696) and a Postdoctoral Fellowship from the Institute for Research in Cognitive Science at the University of Pennsylvania to Toben H. Mintz, a NIH research grant (DC00167) to Elissa L. Newport, and a NIH National Research Service Award institutional training grant (DC00035) to the Center for the Sciences of Language at the University of Rochester. Manuscript preparation was supported in part by an equipment grant to the first author from the Intel Corporation.

References

Bates, E., & MacWhinney, B. (1979). The functionalist approach to the acquisition of grammar. In E. Ochs and B. Schieffelin (Eds.), *Developmental pragmatics*. New York, NY: Academic Press.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the state of the art*. Cambridge: Cambridge University Press.

Bever, T. G. (1992). The demons and the beast: Modular and nodular kinds of knowledge. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing*. Hove, England: Lawrence Erlbaum Associates.

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

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

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

Brill, E. (1991). Discovering the lexical features of a language. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

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

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

Chomsky, N. (1955). *The logical structure of linguistic theory*. New York: Plenum Press.

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

Echols, C. H. (1993). A perceptually-based model of childrens' earliest productions. *Cognition*, *46*, 245–298.

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

Finch, S. P., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentences. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society of America*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gerken, L., Landau, B., & Remez, R. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology*, *26*, 204–216.

Gillette, J., Gleitman, L., Gleitman, H., & Lederer, A. (1999). Human simulations of lexical acquisition. *Cognition*, *73*(2), 135–176.

Gleitman, L. R. (1990). The structural sources of verb meaning. *Language Acquisition*, *1*, 3–55.

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 state of the art*. Cambridge: Cambridge University Press.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

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

Höhle, B., & Weissenborn, J. (1998). Sensitivity to closed-class elements in preverbal children. In A. Greenhill, M. Hughes, H. Littlefield, & H. Walsh (Eds.), *Proceedings of the 22nd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Jelinek, E. (1995). Quantification in straits Salish. In E. Bach, E. Jelinek, A. Kratzer, & B. Partee (Eds.), *Quantification in natural languages* (Vol. II). Dordrecht: Kluwer Academic Publishers.

Johnson, J., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Juliano, C., & Bever, T. G. (1990). *Clever moms: Regularities in motherese that prove useful in parsing*. Paper presented at the 1990 CUNY Sentence Processing Conference.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, *24*, 252–293.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*(1), 15–47.

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

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

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

McNeill, D. (1966). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The genesis of language: A psycholinguistic approach*. Cambridge, MA: MIT Press.

Mintz, T. H. (1996). *The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories*. Unpublished Doctoral Dissertation, University of Rochester.

Mintz, T. H. (2000). Unique entropy as a model of linguistic classification. *Proceedings of the Twenty-second Annual Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mintz, T. H. (2002). Unique entropy, hierarchical clustering, and linguistic classification. Submitted for publication.

Mintz, T. H., Newport, E. L., & Bever, T. G. (1995). Distributional regularities of grammatical categories in speech to infants. In J. Beckman (Ed.), *Proceedings of the North East Linguistics Society 25* (Vol. 2). Amherst, MA: GLSA.

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

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

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11–28.

Newport, E. L., Gleitman, L. R., & Gleitman, H. (1977). Motherese: The speech of mothers to young children. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum Associates.

Peters, A. M. (1977). Language learning strategies: Does the whole equal the sum of the parts? *Language*, *53*, 560–573.

Pinker, S. (1979). Formal models of language acquisition. *Cognition*, *7*, 217–283.

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

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 435–469.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In K. E. Nelson (Ed.), *Children's language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Santelmann, L., & Jusczyk, P. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, *69*, 105–134.

Schlesinger, I. M. (1974). Relational concepts underlying language acquisition. In R. L. Schiefelbusch & L. Lloyd (Eds.), *Language perspectives: Acquisition, retardation, and intervention*. Baltimore, MD: University Park Press.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

Shady, M. E. (1996). *Infants' sensitivity to function morphemes*. Unpublished Ph.D. dissertation, State University of New York at Buffalo, Buffalo, NY.

Shady, M., Gerken, L. A., & Jusczyk, P. (1995). Some evidence of sensitivity to prosody and word order in ten-month-olds. In D. MacLaughlin & S. McEwan (Eds.), *Proceedings of the 19th Boston University Conference on Language Development* (Vol. 2). Sommerville, MA: Cascadilla Press.

Shafer, V., Shucard, D., Shucard, J., & Gerken, L. A. (1998). 'The' and the brain: An electrophysiological study of infants' sensitivity of English function morphemes. *Journal of Speech-Language and Hearing Research*, *41*, 874–886.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*, B11–B21.

Shipley, E. F., Smith, C. S., & Gleitman, L. R. (1969). A study in the acquisition of language: Free responses to commands. *Language*, *45*, 322–342.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103–114.