COGNITIVE SCIENCE Vol 22 (4) 1998, pp. 425–469 ISSN 0364-0213

Distributional Information: A Powerful Cue for Acquiring Syntactic Categories

MARTIN REDINGTON
University College London

NICK CHATER
University of Warwick

STEVEN FINCH
Thomson Technology

Many theorists have dismissed a priori the idea that distributional information could play a significant role in syntactic category acquisition. We demonstrate empirically that such information provides a powerful cue to syntactic category membership, which can be exploited by a variety of simple, psychologically plausible mechanisms. We present a range of results using a large corpus of child-directed speech and explore their psychological implications. While our results show that a considerable amount of information concerning the syntactic categories can be obtained from distributional information alone, we stress that many other sources of information may also be potential contributors to the identification of syntactic classes.

I. INTRODUCTION

In the first years of life, human infants routinely acquire language from a relatively noisy and partial body of evidence. Yet, from a computational point of view, there has been little progress in explaining how this feat can be accomplished, even in principle. Furthermore, there has been relatively little constraint offered by the empirical data on child language (Ingram, 1989), partly because many theoretically relevant manipulations on the child's linguistic input or maturation are thankfully prohibited on ethical grounds. The computational problems involved in acquiring many aspects of language from realistic linguistic

Direct all correspondence to: Martin Redington, Department of Psychology, University College London, London, UK, WC1E 6BT; E-Mail: m.redington@ucl.ac.uk; Nick Chater, Psychology Department, University of Warwick, Coventry, CV4 7AL, UK; Nick.Chater@warwick.ac.uk.

input are indeed formidable, and have led many to argue that the majority of linguistic knowledge must be innate (e.g., Chomsky, 1965). Nonetheless, it may be that progress can be made on providing computational models of certain constrained aspects of the language acquisition problem.

One problem that seems particularly tractable is modeling how the child acquires syntactic categories. We show that a surprisingly simple distributional analysis can be highly informative of syntactic category membership, using a corpus of adult speech taken from the CHILDES project (MacWhinney, 1989; MacWhinney & Snow, 1985). We present a range of results and explore their implications for psychological theories of language acquisition.

These results show that simple distributional evidence is a potentially important source of information for identifying the syntactic categories of words, although we stress that a variety of other sources may also be highly informative. Thus, despite some influential a priori arguments to the contrary (e.g., Pinker, 1984), distributional information does provide a powerful cue for acquiring syntactic categories.

We begin by introducing the problem of learning words' syntactic categories, and then consider the range of possible sources of information that could usefully be employed in solving this problem. We consider the difficulties involved in assessing the potential contributions from each of these sources, and argue that distributional sources, whilst enjoying no theoretical primacy, are methodologically the easiest to investigate and assess. After outlining past distributional approaches within cognitive science and applied natural language processing, we present our method and report its application to the CHILDES corpus, and the results obtained.

The Problem of Learning Syntactic Categories

Acquiring language involves classifying lexical items into syntactic categories. This is a difficult problem from both the nativist and empiricist perspectives on language acquisition.

For the strong nativist, the grammatical rules, including schematic syntactic categories, are innate and the learner's problem is to map the lexicon of the target language into these categories. Clearly, there must be significant constraints on which mappings are considered. A completely unconstrained search with n items and m syntactic categories (assuming, for simplicity, that each item has a single syntactic category), would involve considering m^n possible mappings. With just 20 items and 2 syntactic categories, there are already more than a million permutations. For the empiricist, the search appears more difficult still, since even the number of syntactic categories is not known a priori.

On both nativist and empiricist views, the learner must make the first steps in acquiring syntactic categories without being able to apply constraints from knowledge of the grammar. For the empiricist, this information is simply not initially available. For the nativist, grammatical information initially provides little constraint, since grammatical rules specify possible relations between words under their syntactic description; this description is hard to apply before at least an approximate solution to the mapping problem has been found. Thus, irrespective of the role of innate knowledge of grammar in language acquisition generally, any clues to syntactic category that can be obtained from linguistic and extra-linguistic environmental input would appear to make this aspect of the acquisition problem more straightforward.

What Information is Available?

There are four main sources of information in linguistic input which have been proposed as potentially useful in learning syntax, and which, in particular, may be useful in learning syntactic categories. These are based on distributional analysis of linguistic input; on relating the linguistic input to the situation or communicative context in which it occurs; on phonological cues to syntactic category; and on the analysis of prosody. A fifth source of information, internal to the learner, is innate knowledge of syntactic categories (as opposed to innate knowledge of grammar per se).

Distributional Information

Distributional information refers to information about linguistic contexts in which a word occurs.[1] Various authors (e.g., Finch & Chater, 1992; Kiss, 1973; Maratsos, 1979, 1988; Maratsos & Chalkley, 1980; Rosenfeld, Huang, & Schneider, 1969) have suggested that the fact that words of the same category tend to have a large number of distributional regularities in common can be used as a cue to syntactic category. For example, Maratsos and Chalkley (1980) noted that word roots which take the suffix -ed typically also take the suffix -s, and are verbs. Words which take the suffix -s, but not the suffix -ed, are typically count-nouns. The proposal is that patterns of correlation between simple properties of word roots can therefore be used to infer proto-word classes which can later be refined to the adult word classes. Various other approaches, based on measuring local statistics of large corpora of language have also been proposed (e.g., Brill, Magerman, Marcus, & Santorini, 1990; Finch & Chater, 1991, 1992, 1993, 1994; Marcus, 1991; Schutze, 1993), and we shall consider these further below.

Simple distributional methods are sometimes associated with a general empiricist *tabula rasa* approach to language learning, which has been widely criticized (e.g., Chomsky, 1959; Pinker, 1984). However, this is not germane in the present context, since distributional methods are not proposed as a general solution to the problem of language learning, but rather as a possible source of information about syntactic structure. Furthermore, it may be that there are innate constraints on the possible distributional analyses and learning mechanisms which the learner can apply, and it is possible, though not necessary, that these learning mechanisms might be specific to language. So distributional methods could themselves, in a sense, embody innate knowledge.

Semantic Bootstrapping

Grimshaw (1981), Pinker (1984) and Schlesinger (1981, 1988), though from different perspectives, propose that the mechanism for the initial classification of words makes use of a correlation between prior semantic categories (such as object and action) in terms of which the child already perceives the world and syntactic categories (such as noun and verb). This means that the language learner can use knowledge gained about word meaning as a basis for an initial classification of words. This provides a starting point for other language acquisition processes which ultimately lead to adult categories (see Pinker, 1984 for a sketch of an algorithm; Schlesinger [1981, 1988] for a general description of a different account). On the view that a set of abstract syntactic categories is innately specified, the learner is viewed as making a tentative mapping from lexical items to these syntactic categories, using semantic information (Pinker, 1984).[2]

A somewhat different approach, which also stresses the importance of extralinguistic context, is the "social/interaction" model (see Bruner, 1975; Nelson, 1977; Snow, 1972, 1988, for a range of views). This approach stresses the child's communicative intent and the importance of the development of appropriate communicative relationships with caregivers. The pragmatic purpose to which language can be put by the learner, or by caregivers, is thought to crucially affect the course of acquisition. Thus the correlations between pragmatic referents such as *force of request, object under consideration, and location of object* and syntactic categories such as *verb* and *noun* and *preposition* are exploited to form initial categories (e.g., Ninio & Snow, 1988).

Phonological Constraints

Kelly (1992) has proposed that the many regularities between the phonology of words and their syntactic categories can be used to acquire these categories. English disyllabic nouns, for example, tend to have stress on the initial syllable, while verbs have final syllable stress (e.g., Liberman & Prince, 1977); English polysyllabic words are predominantly nouns (Cassidy & Kelly, 1991); English open-class words are generally stressed more strongly than closed-class words (Gleitman, Gleitman, Landau & Wanner, 1988). Word duration also appears to be a valuable clue (e.g., Sorenson, Cooper & Paccia, 1978). For example, English words occurring clause or phrase finally are typically longer in duration, due to lengthening affects associated with such boundaries (Lehiste, 1970). Since, in English, nouns are more likely to occur in these positions, duration can therefore serve as a cue to syntactic category (Davis, Morris & Kelly, 1992). These and many other cues, both in English and across languages, have been largely neglected in the language acquisition iterature, although they may potentially be exploited to provide useful information concerning syntactic categories (see Kelly, 1992, for a survey of potential relationships between phonology and syntax). Of course, as is the case for all cues which are not universally applicable, some account of how the learner could ascertain which cues are relevant to their particular language will be required.

Prosodic Information

Prosodic contours provide another possible source of constraint. Morgan and Newport (1981) and Hirsh-Pasek et al. (1987) propose that learners exploit the mutual predictability between the syntactic phrasing of a sentence, and the way it is said (i.e., its prosodic phrasing). Consequently if the child takes note of how something is said, he or she has information about the "hidden" syntactic phrasing of the sentence. This information might provide clues about the syntactic properties of words in the input, and thereby constraints on their possible syntactic categories.

Innate Knowledge

On all views apart from tabula rasa empiricism, innate knowledge can bear on the problem of syntactic category acquisition in two ways. Firstly, learning mechanisms which exploit information of any kind in the input may be innately specified or constrained. For instance, the language-internal relationships considered by a distributional method, or the relationship between semantic features of a word (e.g., naming an object) and syntactic ones (e.g., being a noun) may be innately specified. Secondly, innate knowledge or constraints may specify, for instance, the number of syntactic categories, or the relationships between them (for instance that closed class are fewer in number, but more frequent, whilst the converse is true for open class words). Innate knowledge of either kind serves to constrain the search space of the learner.

Assessing the Utility of Information Sources

In order to quantitatively assess the amount of information that could be gleaned by the language learner from each of these sources, it is useful to study the linguistic (and, for semantic approaches, extralinguistic) input actually received by the language learner. Looking at the structure of this input is important because some cues may seem to be very informative, but in fact occur very rarely, while other cues, which may seem unpromising on theoretical grounds, may in practice cooccur reliably with important aspects of syntactic structure.

It is difficult to assess the potential contribution of semantic factors in a quantitative fashion, since it is extremely labor intensive to record the extralinguistic context associated with even a small amount of linguistic input, and furthermore, it is difficult to know what description of that context is likely to be relevant, given the general cognitive apparatus of the language learner.

Prosodic information, since it is internal to the speech stream, may be more easily recorded, but is still labor-intensive to notate. There are currently no large (millions of words) corpora of conversation with detailed prosodic markings. In the future, however, if such corpora are developed, it may be possible to give a quantitative assessment of the amount of information that prosody could potentially give the language learner. Currently, as Jusczyk (1993) notes, the potential utility of prosodic information remains to be determined.

Assessing the utility of phonological cues is reasonably tractable. The value of a number of phonological cues as indicators of syntactic category at the level of individual lexical items for English have been studied (Cassidy & Kelly, 1991; Kelly & Bock, 1988), and some work has been carried out for other languages (Kelly, 1992, notes, for example, studies on French [Tucker, Lambert, Rigault & Segalowitz, 1968], Hebrew [Levy, 1983] and Russian [Popova, 1973]). A wide range of studies of phonologically transcribed corpora from a variety of languages would appear both feasible and potentially illuminating (see Shillcock, Lindsey, Levy & Chater, 1992; Cairns, Shillcock, Chater & Levy, 1995, for related work). There is also potential for computational accounts of how this information might be utilized.

Distributional methods can often be readily explored in practice. In particular, unlike semantic and prosodic approaches, distributional analysis can be conducted over electronically stored texts, represented purely as sequences of distinct words, and these are (at least for English) available to researchers in almost unlimited supply. In addition, reasonably large corpora of transcribed speech, such as the Lund corpus (Svartvik & Quirk, 1980) and the CHILDES database (MacWhinney & Snow, 1985) are available. These are large enough to provide at least some validation of the performance of distributional methods which have been primarily developed using text corpora.

It seems entirely likely that many different sources (including semantic, phonological and prosodic and innate knowledge) may be (perhaps highly) informative about syntactic structure and that the child may draw on them. Additionally, multiple sources of information, and their interactions may be of crucial importance in the acquisition of syntactic categories.[3] However, in view of the methodological considerations described above, such questions are very hard to investigate and we therefore restrict ourselves for the moment to consideration of the potential contribution of distributional methods in isolation.

Is the Study of Distributional Information Useful?

The usefulness of studying the potential role of distributional information in acquiring syntactic categories can be criticized from two opposing points of view.

The first point of view is that the usefulness of such information is obvious. In traditional linguistics, syntactic categories are operationally defined in terms of "distributional tests," which assume that words and phrases with similar distributions are in the same linguistic category. Probably the best known test is the "replacement test":

> "Does a word or phrase have the same distribution (i.e., can it be replaced by) a word or phrase of a known type? If so, then it is a word or phrase of that type" (Radford, 1988).

It seems hardly surprising that distributional information is informative about syntactic categories, because syntactic categories are defined in terms of their distribution.[4]

This argument is incorrect, because it does not recognize the difference between the nature of the distributional information used by linguistics, and the distributional information available to the child. In linguistics, distributional tests involve directly eliciting native speaker intuitions concerning the grammaticality of sentences that the linguist believes to be of particular importance. However, the child cannot elicit judgments, but must observe a noisy and very partial corpus of occurrences of words in a limited range of specific contexts. Moreover, the linguist typically assumes that the category of all other words used in test sentences to be fixed and known, whereas the child must initially begin assigning syntactic categories to words with no prior knowledge of the syntactic category of *any* word. Indeed, the problem for the child is so much harder that it has led theorists to propose the second objection: That distributional information cannot provide any useful information for acquiring syntactic categories.

Perhaps the most influential attack on the usefulness of distributional information is provided by Pinker (1984), who presents four lines of argument.[5]

First, Pinker argues that relationships which are apparent to the learner in the surface structure of language cannot be usefully exploited by distributional methods. He claims that the vast number of possible relationships that might be included in a distributional analysis is likely to overwhelm any distributional learning mechanism in a combinatorial explosion.

Second, he claims, easily observable properties of the input are in general linguistically uninformative: "Most linguistically relevant properties are abstract, pertaining to phrase structure configurations, syntactic categories, grammatical relations, ... but these abstract properties are just the ones that the child cannot detect in the input prior to learning ... the properties that the child can detect in the input—such as the serial positions and adjacency and cooccurrence relations among words—are in general linguistically irrelevant", (Pinker, 1984 p. 49-50).

Third, Pinker argues that "even looking for correlations among linguistically relevant properties is unnecessarily wasteful, for not only do languages use only certain properties and not others, they sanction only certain types of correlations among those properties."

Fourth, Pinker proposes that "spurious correlations" will arise in local samples of the input. For example, the child could hear the sentences *John eats meat, John eats slowly*, and *the meat is good* and then conclude that *the slowly is good* is a possible English sentence (Pinker, 1984).

None of these arguments are persuasive. Pinker's first point, the danger of a combinatorial explosion assumes that distributional learning mechanisms will blindly search for relationships between a vast range of properties. While this may be a fair criticism of early, unimplemented distributional proposals (e.g., Maratsos & Chalkley, 1980), the kinds of learning mechanisms that contemporary researchers have considered and implemented tend to focus on highly specific properties of the input. The case studies below indicate that even very simple and easily observable properties (such as cooccurrence statistics) can be highly informative.

Pinker's second point above relies on equivocation over what is meant by "linguistic relevance." Uncontroversially, generative grammar does not capture the structure of language in terms of serial position, adjacency and cooccurrence relations. However, this is not to say that such relations are not linguistically relevant, in that they carry useful information about the structure of language. Indeed, contrary to Pinker's assertion, all three of the examples he gives can provide information about a word's syntactic category, for English at least. The utility of distributional learning mechanisms, as a technique for investigating language acquisition, is that they allow empirical tests of such assertions. As should be clear from the above, a priori intuitions on such matters cannot be trusted.

Pinker's third point starts from reasonable premises. As languages vary in many respects, it seems likely that different learning mechanisms will be recruited, and that their contributions might differ from one language to the next. But this cannot be condemned as "unnecessarily wasteful." Since the child is able to learn any language, but in actual fact generally faces only one, its learning apparatus is "wasteful" by necessity. Even a strict universal grammar account is "wasteful," in that almost all of the space of possible parameter settings will go unused.

Pinker's fourth point may be a fair criticism of early and underspecified distributional proposals. An important aim in the study of distributional learning mechanisms is to avoid such spurious generalizations. The fact that a brittle and extraordinarily naive approach to distributional analysis, which draws conclusions from single examples, falls prey to such errors is not a valid argument against the entire class of distributional approaches. Without consideration and empirical assessment of more sophisticated approaches such objections are premature.

In the light of these strongly contrasting a priori views concerning the utility of distributional information in acquiring syntactic categories, there seems to be a genuine empirical question, which can only be addressed by analyzing whether plausible learning mechanisms can extract useful information about syntactic categories from corpora of child-directed language. Before reporting our own research, we outline previous work concerning distributional learning methods of potential psychological relevance.

Relevant Distributional Approaches

A number of existing methods are concerned with the analysis of distributional information. These originated in linguistics, applied natural language processing and psychologically motivated computational research. We first discuss distributional analysis as a historically important program in linguistics, and then turn to more recent computational work, using neural networks and straightforwardly statistical methods.[6]

Distributional Analysis in Linguistics

The distributional approach to linguistics (e.g., Fries, 1952; Harris, 1954) was the dominant linguistic approach before the Chomskyan revolution. The distribution of a linguistic item was taken to be ".. the sum of all its environments." (Harris, 1954, p. 146). In the purest form, the distributional view saw linguistics as the project of describing the structural relationship between linguistic items and their contexts. This involved, among other things, classifying together linguistic items which can occur in the same environments or contexts.

Distributional linguists were interested in the discovery of language structure from corpora, purely from the point of view of providing a rigorous methodology for field linguistics; they did not consider that this approach might have any relationship to language acquisition in children. Indeed, Harris and others assumed that behaviorist psychology would account for language acquisition and use and that such matters were not the business of linguistics.

It was assumed that linguistic methodology could proceed from the isolation of phonemes, to the uncovering of morphological and thence syntactic structure. These procedures were useful heuristic guides rather than algorithms (though see Harris, 1955). Aside from the degree of formalization, these methods differ from more recent work in several respects. First, they abstracted away from all frequency information—rare contexts were rated equally with common ones, since the goal was to describe the possible structural relationships in language, rather than those which happen to be the most frequent. Second, distributional linguists conceived of language as an external cultural product, and did not consider it in a psychological or computational context. Third, they were unable to test their methods except with very small samples of language, since long and tedious analysis had to be conducted by hand.

After the development of generative grammar, distributional linguistics was justly criticized on a number of grounds (e.g., Chomsky, 1964), including connections with dubious doctrines such as behaviorism and positivism, lack of formal rigor, a failure to properly deal with syntax, and an over-restrictive definition of linguistics, which ruled out semantics, and any psychological aspects of language. We suspect that the bad name of distributional linguistics has led many researchers to discount the possibility that distributional information of any sort can have any bearing on language and language acquisition. As we shall see below, the possibility should not be so readily discounted.

Neural Network Approaches

The most influential neural network approach to learning the structure of sequential material (which here refers to the prediction of the next item in a sequence) uses simple recurrent networks (SRNs) due to Elman (1990, 1991; Cleeremans, Servan-Schreiber & McClelland, 1989). One of the most impressive properties of SRNs is that they appear to assign similar hidden unit patterns to items which have the same syntactic category in a simple grammar. Elman (1990) trained an SRN to predict the next item in an input sequence (the sequence was generated by a grammar containing 29 lexical items, 12 syntactic categories, and 16 rules, which generated two and three word sentences, which were concatenated without punctuation). When the hidden unit activations associated with each item were appropriately averaged and subjected to cluster analysis,[7] the resulting classification reflected many of the distinctions between syntactic category present in the original grammar.

Another approach to learning the linguistic categories of small artificial languages uses a competitive network to produce a topographic mapping between the distribution of contexts in which an item occurs and a 2-dimensional space (Ritter & Kohonen, 1989, 1990; Scholtes, 1991a, 1991b). The results show that items with the same linguistic category tend to lie in neighboring regions of the space.

In order to undertand what aspects of the input these methods are picking up, Chater and Conkey (1992, 1993; see also Conkey, 1991) compared the output of a cluster analysis on hidden unit patterns in an SRN and a cluster analysis of a simple distributional statistic of the training set. They proposed that, since the SRN's goal is prediction, its hidden unit values will reflect the distribution of probabilities associated with each possible next item. Chater and Conkey recorded the number of times each successor followed the target item, in each context, and then averaged across contexts, as before. This is equivalent to simply recording the number of occurrences of each successor to each target item. These sets of probabilities were used in place of the averaged hidden unit patterns, and were clustered as before. The resulting classification was extremely close to that obtained from the SRN analysis. This outcome suggests that the average hidden unit pattern in the SRN associated with a lexical item reflects the distributional statistic measured directly from the training set: that is, the distribution of probabilities of each possible continuation.[8]

Both the SRN and Kohonen network have two limitations:

1. It has not yet been possible scale up from very small artificial data sets to deal with real linguistic data. For example, in SRNs learning becomes extremely inefficient and slow, if it occurs at all, as vocabulary increases and the language become more complex, since prediction becomes more difficult (Chater & Conkey, 1993).
2. The linguistic categories are implicit within these networks, and can only be revealed using a subsequent cluster analysis. Thus, a significant amount of the computational work in approximating syntactic categories is not performed by the network itself.

Scaling is not so problematic, however, if simple distributional statistics are collected directly from the training set. Collecting such statistics requires only a single pass through the training set, and scaling problems which are related to the minimization of prediction error are avoided. In contrast, the SRN required 50 passes through an input of 2,700 items with the simple grammar (Elman, 1990), and a significant proportion of simulations may fail to train successfully (Conkey, 1991).

Statistical Approaches to Language Learning

A number of direct statistical methods have been proposed which relate to the problem of finding syntactic categories. The problem of learning linguistic categories generally is relevant to a number of aspects of practical natural language systems. An early example is Rosenfeld, Huang and Schneider (1969) who applied cluster analysis to simple distributional statistics for small corpora. The upsurge of interest in statistical methods in computational linguistics (Charniak, 1993) has led to a range of approaches using very large text corpora (e.g., Brill, 1991; Church, 1987; Garside, Leech, & Sampson, 1985; Kupiec, 1992; Marcus, 1991; Schutze, 1993). The main aim of these approaches is practical utility, and deriving linguistically coherent categories is not a primary goal. Hence, this work, while suggestive, does not directly explore the extent to which syntactic category information could be derived from distributional evidence in language acquisition. A small number of studies have, however, explicitly aimed at assessing the potential contribution of distributional information.

Kiss (1973) described a complicated model of category acquisition, in neural-network like terms, but implemented purely statistically. He performed cluster analysis on the conditional probabilities of the successors to each target word (this is very similar to the analysis of Chater and Conkey, 1993, above). Using a 15,000 word sample corpus of mother-to-child speech, where the children were between 8 and 35 months in age, and considering only 31 high-to-medium frequency words, the classification resulting from the cluster analysis showed clear groupings for nouns, verbs, and adjectives, whilst three less clear-cut clusters contained prepositions, pronouns, and determiners. Due to both limitations of available corpora, and computational resources, Kiss was unable to extend this promising work.

Wolff (1976, 1977, 1988) has proposed that aspects of acquisition can be modelled by using a distributional analysis to find frequently occurring sequences in the input. Using small artificial grammars (e.g., 2 or 3 word sentences, with a 12 word vocabulary, presented as continuous strings of letters) and very small, simple, natural language texts, he shows that these frequent 'chunks' correspond to linguistically meaningful units. Like the neural network approaches, this method does not appear to scale up readily to more realistic natural language input.

Results from both neural network and statistical distributional analyses have been suggestive but have not demonstrated the utility of distributional information for realistic linguistic input. We show below, however, that this kind of approach can be extended to provide information concerning syntactic categories even for very large and rich corpora, and therefore that distributional analysis is a potentially useful source of information in identifying words' syntactic categories.

A New Distributional Approach

We aim to demonstrate that words' distributional properties can be highly informative of syntactic category and argue that this information can be extracted by some psychologically plausible mechanisms. We propose that using distributional information concerning syntactic categories involves three stages:

1. Measuring the distribution of contexts within which each word occurs.
2. Comparing the distributions of contexts for pairs of words.
3. Grouping together words with similar distributions of contexts.

We shall consider each of these below.

Measuring the Distribution of Each Word

Collecting distributional information involves collecting information about contexts in which words occur. What should count as a "context" for a word? The promising results obtained both by Kiss and SRNs suggest that a useful notion of context may be defined simply in terms of the distribution of words which occur near the target word. Where Kiss used only immediate successors, we consider a range of different contexts below. Broader notions of context have also been used—for example, Lund and Burgess (1996) considered items in a large "window" around the target word, weighted by their nearness to the target (although their method was aimed at providing information relating to semantic rather than syntactic properties of words).

The cooccurrence measures between pairs of items in some fixed relationship that we consider are traditionally known as bigram statistics. These statistics[9] are very straightforward to measure, and are appropriate in the present case because they do not presuppose any knowledge of syntactic structure (which even if possessed, cannot be applied initially, as argued above).

A record of such statistics can be viewed as a contingency table. The rows and columns of the table are indexed respectively by a set of target words (the items whose distributions are being measured) and a set of context words (the items which are considered as context). Each cell of the table records the number of times that the relevant context word cooccurred in the appropriate position (e.g., as the next word) with respect to the target word.

For example, given the input *the cow jumped over the moon*, where *jumped* was the current target word, the cells indexed by (*jumped, the*), (*jumped, cow*), (*jumped, over*), and (*jumped, the*), would be incremented in the contingency tables corresponding to, respectively, the previous but one word, the previous word, the next word, and the next but one word.

It is not necessary (or even desirable) to record these statistics for every word in the input in order to provide useful information. From a psychological perspective, in the early stages of syntactic category acquisition, it seems unlikely that a syntactic category will be assigned to every word in the child's input, particularly given that the child's vocabulary is very limited. It may also be computationally appropriate to focus on a small number of target words in order to provide more reliable distributional information and to avoid unnecessarily complex computation. Moreover, it may be appropriate to be even more restrictive with respect to the set of context words (over which frequency distributions are observed). This is because each target word may occur in a relatively small number of contexts, and only the most frequent words in these contexts will provide reliable frequency information. In the analyses below, we explore the effects of varying the size of the sets of target and context words.

Once the bigram statistics have been collected, the row of the contingency table corresponding to each target word forms a vector representation of the observed distribution of the context words in the relevant position, which we shall term a *context vector*. Where more than one position of context is considered (e.g., the immediately preceding and immediately succeeding words), a representation of the overall observed distribution can be formed simply by stringing together the context vectors from the contingency tables for each position.

Comparing the Distributions of Pairs of Words

The overall context vector for each target word can be thought of as a point in a space of possible distributions of contexts. In line with the replacement test, we might expect words with the same syntactic category to have similar distributions (i.e., to lie close together in the space). In order to exploit any information regarding syntactic categories some measure of similarity between distributions is required.

There are several candidate measures for vector similarity which give results in quite good agreement with standard linguistic intuitions. In the CHILDES experiment that we report below, we use the Spearman rank correlation coefficient, ρ, between the context vectors of target words, which produced the most satisfactory results.

Since the rank correlation between two vectors is in the range [-1,1] and negative distance is meaningless, we used an appropriate rescaling of values into the range [0,1]. We have also used a variety of other measures (Euclidean distance, information-theoretic divergence, Pearson's and Kendall's correlation statistics etc., see Finch [1993] for details).

The Rank correlation measure may be the most successful because it is a robust measure which makes no assumptions about the underlying structure of the set of points in the space (Hettmansperger, 1984). This distribution is non-normal and the absolute differences between points on some dimensions can be very large, which may potentially swamp all other differences if parametric measures are used (e.g., Euclidean distance). In fact, these large differences are inevitable, as bigram frequencies, like word frequencies, have an extremely skewed distribution (specifically, they follow Zipf's [1935] law). Intuitively, in linguistic terms, the distribution is non-normal, since the items tend to be clumped within distinct regions of the space (corresponding, to some extent, to syntactic categories). Again, it is intuitively apparent that some elements of the vectors will be orders of magnitude larger than others, reflecting the fact that some words appear in almost stereotyped relationships (e.g., *of the*, *in the*, *of a*).

Given some means of comparing the similarity of distributions of words (i.e., distances in the space of possible distributions of context) this measure can serve straightforwardly as a cue to whether two words belong to the same syntactic category. The more similar the words'distributions (context vectors), the more likely that they are members of the same category. Although the relationship between distance in the space and syntactic category membership will almost certainly not be perfect, even an imperfect relationship can be exploited (and combined with other cues) in order to provide information about syntactic category membership.

Grouping Together Words with Similar Distributions

According to standard linguistic theory, syntactic categories have rigid boundaries. Therefore, the goal of syntactic category acquisition is to assign words to these discrete categories. This would require some kind of non-hierarchical classification over the similarity space. According to alternative linguistic analyses (e.g., Taylor, 1989), linguistic categories may have a prototype structure. On this view, the goal of acquisition is to decide to what extent each lexical item is "noun-like," "verb-like" and so on. Thus, a discrete partitioning of the lexicon would not be appropriate. Finally, exemplar-based views of cognitive processes would suggest that many linguistic generalizations may be based on the similarity of novel items to "neighboring" items, in a similarity space (e.g., Nakisa & Hahn, 1996). According to this viewpoint, there is no need for any explicit syntactic categories to be formed. We remain neutral with respect to these theoretical viewpoints. However, in order to examine the extent to which clusters of similarly distributed target words do belong to the same syntactic category, some method of identifying such clusters (corresponding to regions in the space) is required. For purposes of assessment, we used a standard hierarchical cluster analysis (Sokal & Sneath, 1963), known as average link clustering. The algorithm starts by combining items which are closest together according to the similarity metric. Once items are combined a "cluster" is formed, which can itself be clustered together with nearby items or other clusters. The distance between two clusters is the mean of the distances between the members of each.

The result of this process is a hierarchical structure, with clusters at various scales. The hierarchical structure can be drawn as a *dendrogram*, branching from left to right (e.g., see

Figure 1 below). Clusters correspond to nodes in the dendrogram, and the tighter the cluster (the more bunched its elements are in the space) the further the corresponding node is from the root of the tree (i.e., the similarity between clusters, according to the chosen metric, increases from left to right). This analysis allows a visual presentation of the similarity structure of the space of observed distributions of contexts. The dendrogram can also be "cut" at a particular level of dissimilarity. This results in a set of discrete categories, each corresponding to one of the nodes immediately below the chosen level, and having as their members the items corresponding to their subordinate nodes. This allows quantitative analysis of the extent to which the similarity structure of the space corresponds to syntactic relationships, at a variety of scales.

The use of average link clustering in particular is not crucial. There are a large number of specific variants of this kind of algorithm (Sokal & Sneath, 1963), many of which may be expected to produce similar results.

II. EXPERIMENTS

In this section, we report a range of computational experiments using this method. We begin by describing the corpus that we used as input to the distributional method, the benchmark classification the syntactic categories of words that we used to evaluate the performance of the method, and the scoring scheme that we used to assess the results of the distributional analysis against the benchmark. We then give an overview of the results obtained using this approach, in qualitative terms, before describing a series of specific experiments exploring psychologically relevant properties of the distributional method.

Corpus

The experiments described below were performed using transcribed speech taken from the CHILDES database (MacWhinney & Snow, 1985). CHILDES is a machine-readable collection of corpora of child and child-related speech, transcribed by a number of investigators, and largely recorded in informal North American domestic settings. We used only the English language transcriptions involving non-impaired speakers. We indexed each utterance by sex and age of speaker, taking this information from the documentation accompanying the transcriptions. The resultant corpus contained several million words of speech, from nearly 6,000 speakers. The analysis described here was conducted on adult speech only. Whilst there is no guarantee that the whole of the adult speech in the corpus was child-directed, it would seem to form a fair representation of the speech to which a young child might be exposed. The adult speech corpus comprised over 2.5 million words, from over 3,000 speakers, roughly 2/3 of whom were female.

We did not clean up, or alter, the corpus in any way, apart from stripping away the CHILDES coding information, capitalization, and punctuation. The resulting corpus retained boundaries between utterances, but each utterance was an unpunctuated list of words. The corpus was rather noisy, with false starts, ungrammatical speech and made-up words. Furthermore, since we deliberately did not preprocess the corpus, different transcriptions of the same word were treated as completely separate words, so that, for instance, *wanna* and *wannaa* were effectively different words, as were *mommy* and *mummy*. This large and noisy corpus of adult speech provides a full-scale and realistic test of the usefulness of distributional information as a potential cue to linguistic categories. Indeed, in some ways, the corpus presents a greater challenge than that faced by children, because the number of speakers, dialects, constructions, topics, and vocabulary items is large. The language to which a single child, interacting with a small number of adults, is exposed will tend to be much more homogeneous.

Benchmark Classification

In order to gain some measure of the information about grammatical categories that is conveyed by the distributional analysis, it is necessary to have some benchmark categorisation for each word. Although many words have more than one syntactic category, the method described here does not distinguish between these, but clusters each word according to its distribution over all its contexts. Thus we chose as the benchmark classification of each target word the syntactic category within which it most commonly occurs. It is important to realise that syntactic ambiguity is a common feature, in English at least, and that at some point, learners of English will have to accommodate to this. However, the current method is not capable of distinguishing between multiple syntactic categories. All target words were assigned their most common syntactic category, based on the classifications obtained from the Collins Cobuild lexical database (which contains frequency counts of words' syntactic category over 15 million words of written English, and 1.5 million words of spoken English). These are not necessarily the most frequent usage within the CHILDES corpus. However, they do provide a gold standard of relatively unambiguous categorizations for each word. Only the major syntactic categories were considered, and these are shown in Table 1. There were no occurrences of simple contractions (e.g., *g'day*) in the sets of words

TABLE 1
The Major Categories from the Collins Cobuild Lexical Database, Together with Examples of Each Category from the Set of Target Words, and the Number of Target Words Assigned to Each Category (On the Basis of their Most Frequent Reading According to the Database)

Category	Example	n
noun	truck, card, hand	407
adjective	little, favorite, white	81
numeral	two, ten, twelve	10
verb	could, hope, empty	239
article	the, a, an	3
pronoun	you, whose, more	52
adverb	rather, always, softly	60
preposition	in, around, between	21
conjunction	cos, while, and	9
interjection	oh, huh, wow	16
simple contraction		0
complex contraction	I'll, can't, there's	58

Note. 44 words remained unclassified.

used in the experiments reported below. The complex constructions consisted of two main types; pronoun + modal verb (e.g., *you're*) and modal verb + negation (e.g., *couldn't*). No distinction was made between these in the database (although some distinction was shown in the dendrogram). Many highly frequent words in the corpus were not listed in the Collins Cobuild database (for example, approximately 10% of the most frequent 1,000 words in the corpus were not listed). These were mainly proper names (which were uncapitalised—the database was case-sensitive), nouns with low frequency in adult language (e.g., *playdough*), alternative transcriptions (e.g., *wannaa*, *hafta*), interjections such as *oop*, and *woops*, or nonwords such as *da*. Obvious proper names, alternative transcriptions, and low-frequency nouns were classified appropriately by hand. Interjections, syntactically ambiguous words (e.g., *christmas*, *french*, *indian*, *tv*) and non-words were all left unclassified. The 44 unclassified target words were eliminated from all further quantitative analysis.

Scoring

In order to obtain a quantitative measure of the degree to which the structure of the space of observed distributions of contexts (as reflected in the dendrogram) agrees with the benchmark classification of the syntactic categories of words, we used the following method. First, we "cut" the dendrogram at a range of levels of dissimilarity, obtaining a grouping of discrete sets of words. At one extreme, where words must be very similar indeed to be assigned to the same group, all clusters correspond to a single word. At the other extreme, where very dissimilar words can be clustered together, then all items are grouped into a single cluster. Clearly, the interesting information in the dendrogram is revealed when the dendrogram is cut at intermediate levels of dissimilarity, where words with similar distributions (to some degree) are grouped together, but words with dissimilar distributions are kept apart.

We can evaluate the degree to which the benchmark syntactic categories are reflected in the dendrogram by considering how much the various groups obtained by cutting the dendrogram agree with the benchmark. To do this, we require some way of "scoring" the degree to which two groupings of the same items are in agreement. We used two different kinds of scoring. The first, derived from signal detection theory, is becoming standard in this area (Brent & Cartwright, 1997; Christiansen, Allen & Seidenberg, in press). This consists of two measures: One measure, *accuracy*, is the proportion of pairs of items which are grouped together in the derived groups which are also grouped together in the benchmark groups. The other measure, *completeness*, is the proportion of pairs of items which are grouped by the benchmark which are also grouped together in the derived groupings.

In the language of signal detection theory, a "hit" is a case when the derived grouping correctly clusters a pair of words together (correctly in that they belong to the same category according to the benchmark), a "miss" is where the derived grouping separates a pair of words that should be clustered together, and a false alarm is where the derived grouping clusters together a pair of words that belong to different benchmark categories. Accuracy and completeness can be calculated straightforwardly via the following equations:

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}.$$

$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}}.$$

To understand the significance of accuracy and completeness, consider the following extreme cases. When all items are grouped into a single cluster, then 100% completeness is achieved, because every pair of items that the benchmark classifies together are also grouped together by the derived grouping. However, accuracy is extremely low because although the pairs clustered by the benchmark are grouped together in the derived grouping, so are all the other pairs, which the benchmark treats as distinct. Conversely, consider the case where all groups consist of a single item, except one, which contains a pair of items, which are in the same category according to the benchmark. For this grouping, accuracy is 100%, because the only pair of items that the derived grouping clusters together is correct according to the benchmark. Completeness however, is very low, because all the other pairs which are grouped together by the benchmark are not grouped together in the derived grouping. Notice that overall performance level depends on how well these goals can be achieved simultaneously.

The second kind of scoring that we used was information-theoretic. The signal-detection based scoring described above yields two measures, which intuitively capture the goodness of a classification with respect to the benchmark. However, having two measures for each classification makes comparing the scores for two different classifications difficult. Specifically, when one classification has a higher accuracy, but a lower completeness than a second, or vice versa, it is unclear how accuracy and completeness should be traded off. This second information-theoretic kind of scoring produces only one measure, avoiding this problem. The measure of goodness is the mutual information between the classification and the benchmark (the information that they share), as a percentage of their joint information (the information conveyed by the classification and the benchmark together). This measure reflects both accuracy and completeness. Groupings in either classification or benchmark that are not reflected in the other (that is, both false alarms and misses) will increase the joint information, and penalize the measure. This measure, which we shall term informativeness, is given by:

$$\text{Informativeness} = \frac{I_i + I_j - I_{ij}}{I_{ij}}.$$

where I_i is the amount of information in the classification, as given by:

$$I_i = -\sum_i p(i) \log_2 p(i).$$

where $p(i)$ is the probability that an item is clustered in cluster i. I_j is calculated similarly, over the benchmark categories. I_{ij} is given by:

$$I_{ij} = -\sum_{ij} p(ij)\log_2 p(ij).$$

where $p(ij)$ is the probability that an item occurs in cluster i and benchmark category j.

To understand how this measure relates to accuracy and completeness, consider again the extreme cases described above. When all items are grouped into a single cluster, I_i is effectively zero, and $I_{ij} = I_j$, so that informativeness is also zero. Conversely, when all clusters consist of a single item except one, which consists of a correctly clustered pair of items, I_i will be high, but I_{ij} will still be almost equal to $I_j + I_j$, so that informativeness, while not zero, will have a relatively low value. Generally, both hits (correctly grouped pairs of items) and correct rejections (appropriately separating items of different kinds) will increase mutual information ($I_i + I_j - I_{ij}$), and increase informativeness. False alarms and misses will increase the amount of joint information (I_{ij}) and decrease informativeness.

In the results presented below, we report accuracy, completeness, and when comparing classifications, informativeness, for clusterings at various levels of similarity. The method we describe does not provide any means of selecting an optimum level at which to cut the hierarchy into discrete categories. We do not regard this as a major obstacle, because we view distributional information as one of many sources of information that will contribute to categorization (which, as discussed above, need not necessarily result in discrete categories). As long as distances in the space of observed distributions of context carry information about syntactic relationships, then this distance information can be usefully exploited by the mechanism responsible for category formation. The category formation mechanism is likely to utilise the agreement between distributional information and other sources in order to decide how much to rely on distributional information, and the level of similarity at which it is appropriate to partition the space of observed distributions of context. Below we report results for accuracy, completeness, and informativeness at "optimum" levels of similarity, where these levels were chosen by hand. These results represent a quantitative upper bound on the utility of distributional information obtained from the current method.

In order to show that the method genuinely does provide useful information, we compare the method's performance against a random baseline. For each level of similarity, we held the number of derived clusters, and the number of members of each cluster constant, but randomly assigned items to derived clusters, and then calculated accuracy, completeness, and informativeness as above. This means that random baselines must be calculated afresh for each different analysis. Moreover, we shall see below that the performance of random baselines may differ considerably between analyses, because these analyses differ concerning the number and size of clusters at each level of the dendrogram. In principle, it would be appropriate to conduct statistical tests to determine whether results reliably differed from baselines. However, in practice this is not necessary, because the result of the distributional analysis is always very many standard deviations above baseline—indeed, the standard deviations of the baseline results are so small that they could not be shown clearly in the figures presented below (e.g., Figure 7).

Qualitative Description of Results

Before describing our quantitative results, we first report the performance of the method in qualitative terms. Below we shall show the effect of varying a range of parameters in the method, but here we simply show some qualitative results using parameter values which have been shown to be work reasonably well from our previous experimentation. However the results shown here are reasonably typical of those obtained using this family of methods.

Specifically, we used the most frequent 1,000 words as the target words (the items to be classified), and the most frequent 150 words as context words (the items over which distributional statistics are recorded). Context vectors for the next word, the next but one word, the previous word and the previous but one word, were constructed from the entire 2.5 million words of the CHILDES corpus. The four 150 dimensional vectors for each target word were strung together into a single 600 dimensional vector, which were compared using Spearman's rank correlation as outlined above.

Figure 1 shows the categories resulting when the dendrogram is cut at level 0.8. At this level, the dendrogram consists of 37 discrete clusters. However, the majority of these contain very few items—the 12 clusters that contain 10 or more words contain 910 words in total. The 25 small clusters generally merge appropriately into the larger clusters at a higher level of dissimilarity. However, space precludes showing all of the clusters here.

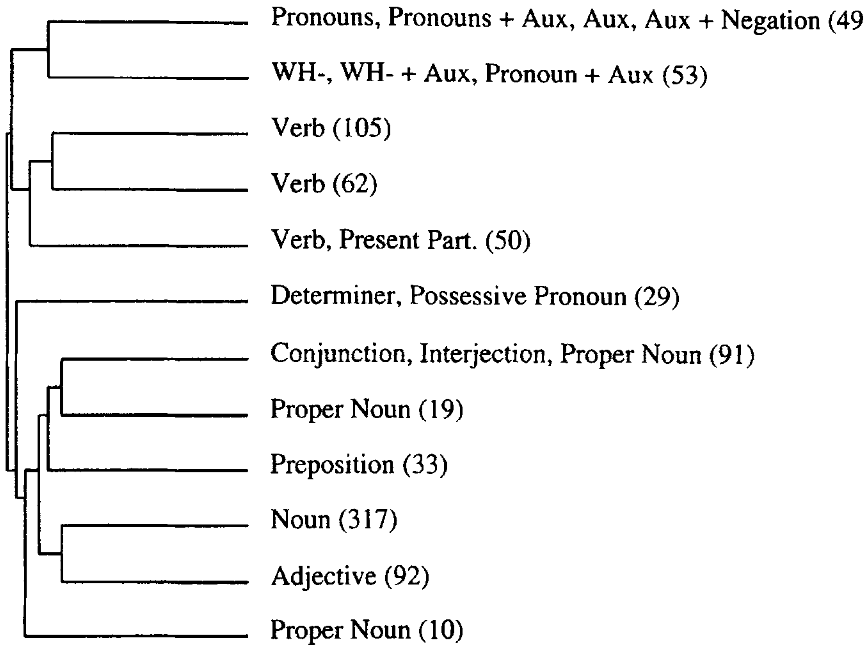
Pronouns, Pronouns + Aux, Aux, Aux + Negation (49
WH-, WH- + Aux, Pronoun + Aux (53)
Verb (105)
Verb (62)
Verb, Present Part. (50)
Determiner, Possessive Pronoun (29)
Conjunction, Interjection, Proper Noun (91)
Proper Noun (19)
Preposition (33)
Noun (317)
Adjective (92)
Proper Noun (10)

Figure 1. The discrete clusters at a similarity level of 0.8 from the analysis of the CHILDES corpus. The clusters have been labelled by hand with the syntactic categories to which they correspond. The number of items in each cluster is shown in parentheses. Only clusters with 10 or more members are shown here.

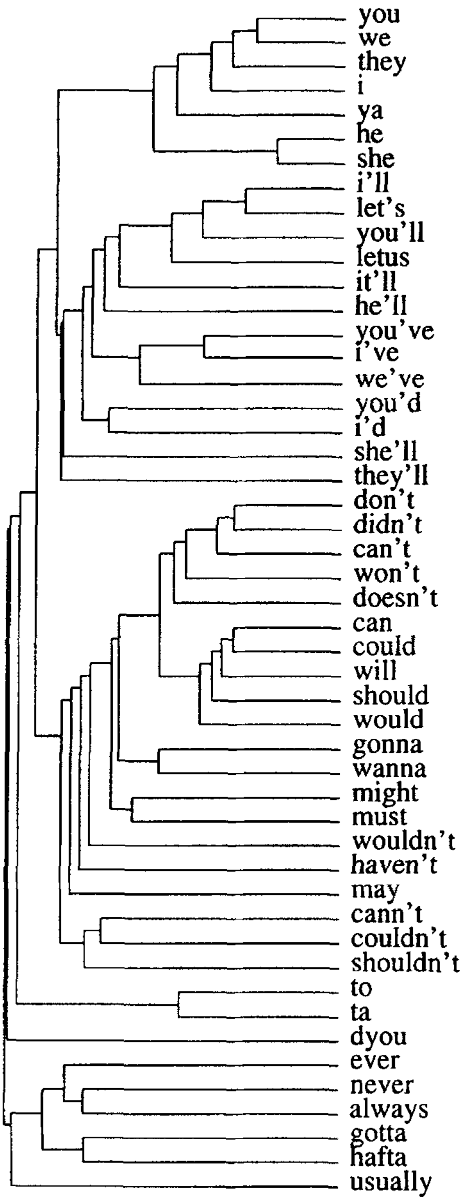

Figure 2. The cluster corresponding to "pronouns, pronouns + auxiliary, auxiliary, and auxiliary + negation" in Figure 1

The labels in Figure 1 were chosen by us to convey some idea of the content of each cluster. In order to illustrate the coherence of these clusters, we show a few selected (but typical) samples here. Figure 2 shows the cluster corresponding to "pronouns, pronouns + auxiliary verb, auxiliary verb, and auxiliary + negation" in Figure 1. It can be seen that whilst some linguistically unrelated items intrude, the pattern of clustering intuitively captures some syntactic relationships. Figure 3 shows the clusters of 62 verbs and the cluster

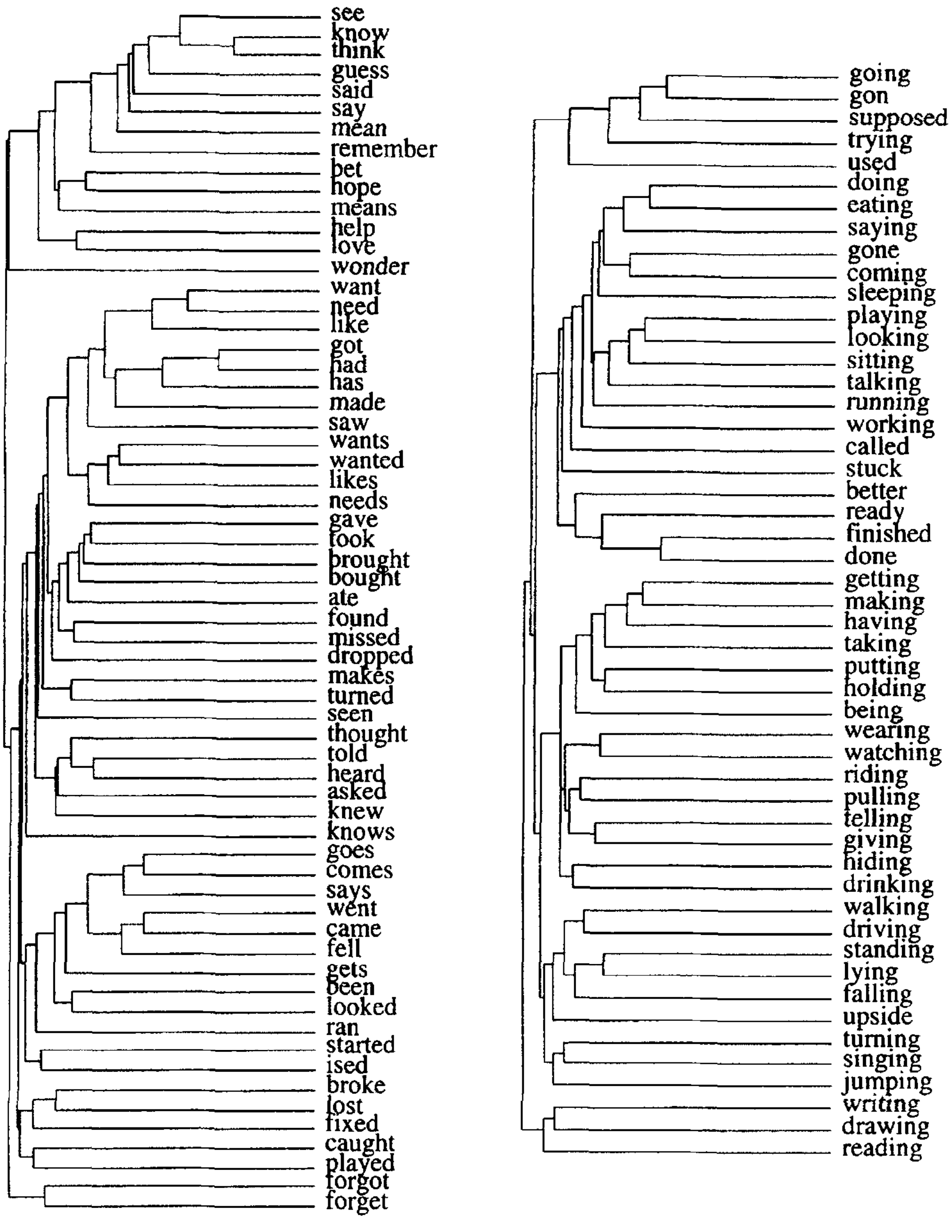

Figure 3. The 62 word "verb" cluster, and the 50 word "present participles" cluster from Figure 1

of 50 present participles from Figure 1. These verb clusters tend to be very coherent—this is equally true of the cluster of 105 verbs that we don't show here. Figure 4 shows the cluster of "conjunctions, interjections, and proper nouns" from Figure 1. It is unclear why a separate cluster of proper nouns is also found elsewhere in the overall dendrogram. Given the clustering of the names shown in Figure 4 it seems possible that these particular names

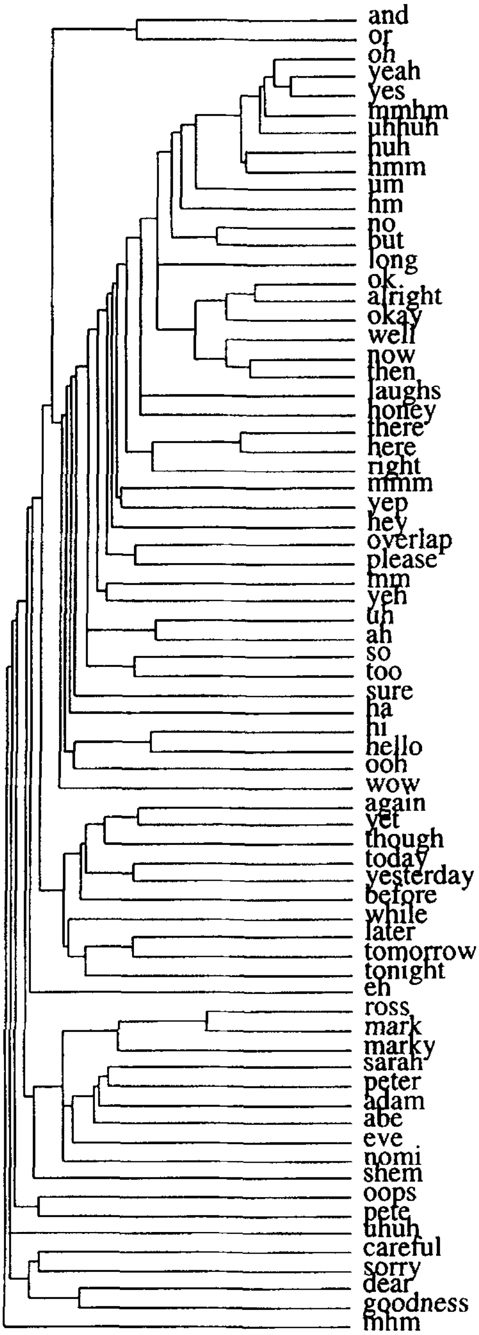

Figure 4. The cluster of "conjunctions, interjections, and proper nouns" from Figure 1

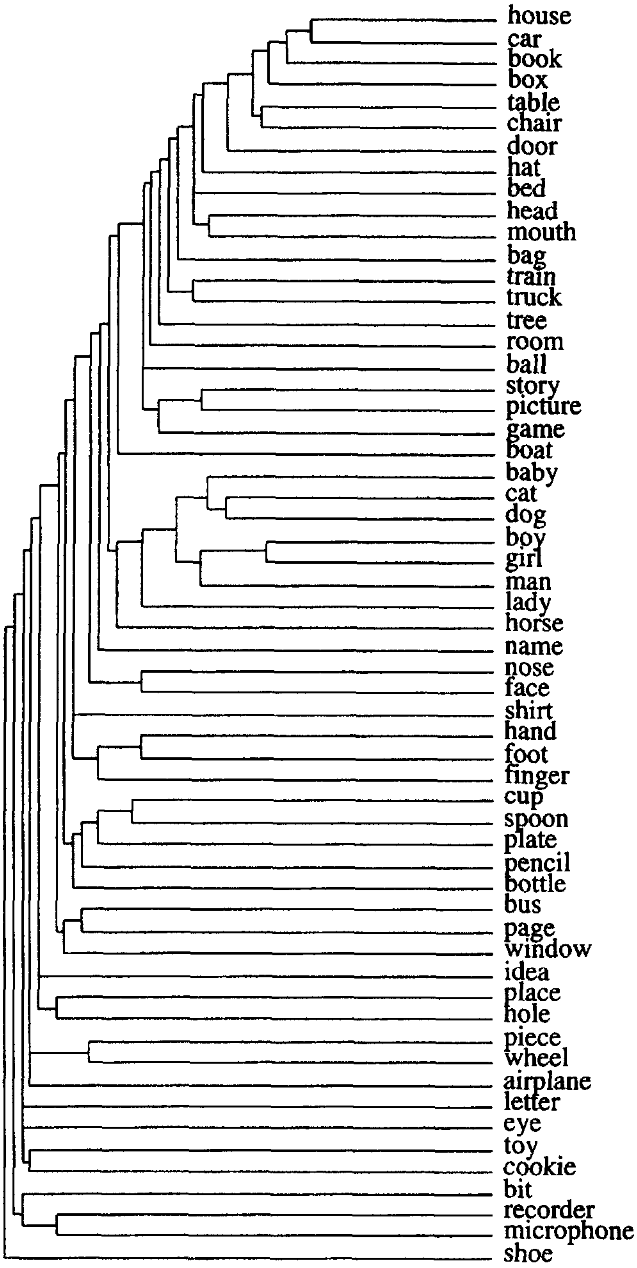

Figure 5. A subcluster of the "Nouns" cluster from Figure 1

(which typically refer to individual children, such as the well-known Adam and Eve), often occurred as single word utterances. Figure 5 shows a subcluster of the 317 word noun cluster from Figure 1. Like the verb cluster, the noun grouping is highly coherent. This cluster also shows clear evidence that the method captures some semantic relationships, most

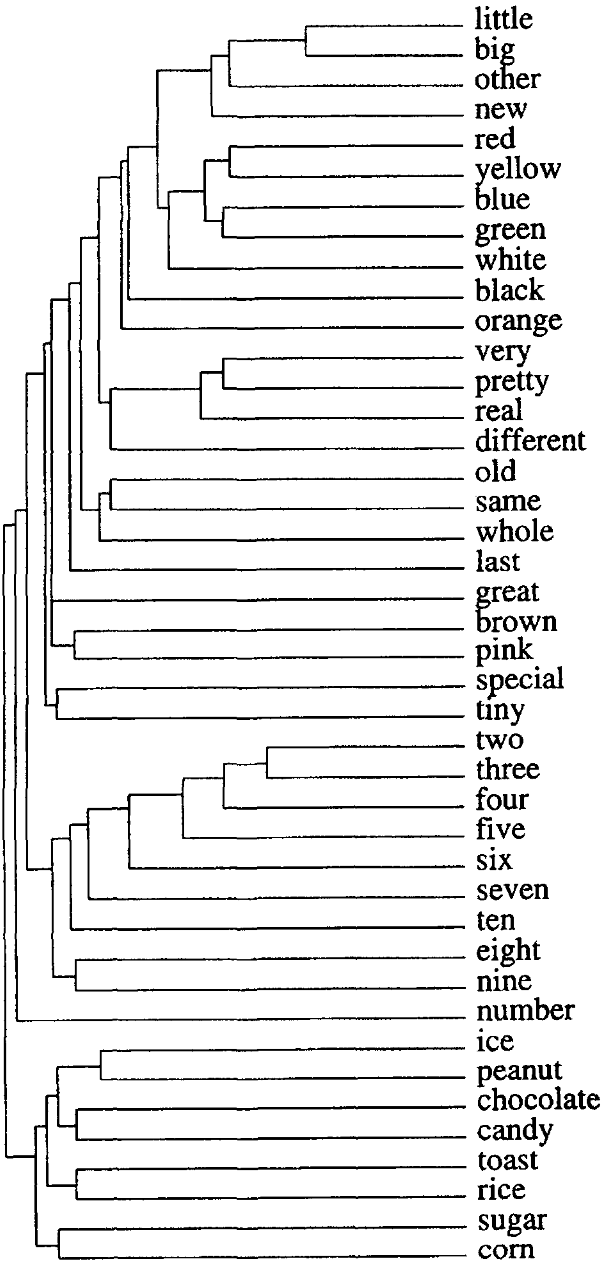

Figure 6. The "adjectives" cluster from Figure 1

clearly in the clustering of food related items. Finally, Figure 6 shows a cluster of adjectives, again with some degree of semantic relatedness evident with respect to color, number and food related adjectives. Although many of these food related adjectives are typically thought of as nouns, they served as adjectives (or as the first half of compound nouns) in the corpus (e.g., *peanut* butter).

These results clearly illustrate that in a qualitative sense, distributional information provides a valuable cue to syntactic category membership. Below we report a number of quantitative experiments, investigating psychologically relevant manipulations of the input and parameters of the method. This allows us to better assess the psychological viability of the claim that children exploit distributional information in acquiring syntactic categories.

Experiment 1: Different Contexts

We have shown qualitatively that information about the distribution of contexts is informative about the syntactic category of lexical items. In order to exploit this information, the child must be able to detect the relevant relationships in the input. Therefore, psychologically important questions are which distributional relationships provide useful information about syntactic categories, and can these relationships be easily detected by the child? The above results were obtained using a particular set of relationships between the target word and a context consisting of the two words to either side of the target. We now consider the extent to which different notions of context are informative about syntactic categories. We investigate whether "preceding" context is more or less informative than "succeeding" context, and how the informativeness of context words changes with distance from the target word, in order to determine which contextual information would be useful to the child.

We first assessed the informativeness of individual context items (i.e., positions in relation to the target word). Figure 7 (A, B, C, and D) shows results of our quantitative analysis with the first, second, third, and fourth succeeding context positions. Each figure shows accuracy and completeness for one context position, where both accuracy and completeness are compared with random baseline values (as discussed above). Specifically, the random baseline scores are average of 10 random allocations of items to categories. As mentioned above, the standard deviations of the results from these random allocations are too small to be shown clearly in Figure 7.

The overall pattern that emerges from Figure 7 is that the nearer the context position to the target word, the more information it carries about the syntactic category of the target. Thus, the context position immediately succeeding the target (Figure 7A) clearly provides some useful information: When the dendrogram is cut at the 0.8 level of similarity, accuracy is .42 and completeness .58, while the random baseline scores are .26 and .36 respectively. However, the context positions which are more distant from the target word (Figure 7B-7D) are much less informative.

Looking at the results for the immediately succeeding word (Figure 7A) in more detail, note that accuracy does not diminish monotonically as the level of similarity increases. Rather, the accuracy curve is "humped." This occurs because within the space of possible distributions of context the words of the same syntactic category are clumped. As similarity increases to the point where these clumps emerge as discrete clusters, accuracy receives a large boost. This humped accuracy curve is a common feature of any of the analyses reported below, and can be interpreted as indicating the level of the dendrogram at which the correspondence between the clusters and syntactic categories is relatively good.[10]

Figure 8 (A, B, C, and D) shows results of our quantitative analysis with the first, second, third, and fourth preceding context positions. In general, preceding context appears to

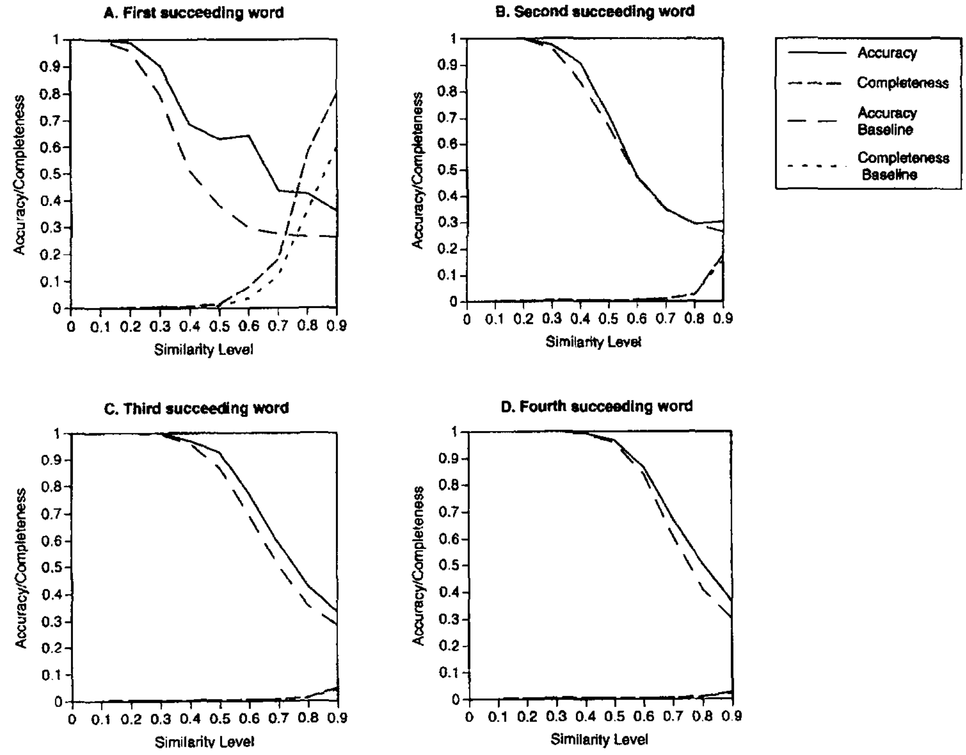

Figure 7. Accuracy and completeness when the first (A), second (B), thhird (C), and fourth (D) succeeding words are used as context. The accuracy curves decrease as similarity increases, while the completeness curves increase. For both accuracy and completeness, the lower of the two lines is the random baseline, averaged over 10 random simulations. Standard deviations were too small to be shown.

be much more useful than succeeding context. This is true for both the preceding word, and the previous word but one. Beyond this window accuracy and completeness approach the random baseline level. As before, the shorter the distance between context position and target word, the more information about syntactic categories, with best results being obtained for the immediately preceding context position.

Having considered each context position in isolation, we next examined the effects of considering multiple context positions. As outlined above, the context vectors for each context position are simply strung together to form a representation of the overall context. The general pattern of results was as follows. When the first and second succeeding context positions were combined, performance was relatively poor, as expected given that the second succeeding context position provides very little information. However, combining the two preceding context positions resulted in considerable improvement in accuracy, at the cost of completeness: At the 0.8 similarity level, accuracy and completeness were 0.72 and 0.47 (with random baselines of .27 and .17). Completeness dropped gradually if wider contexts (i.e., the third and fourth preceding words were included). An intuitive explanation

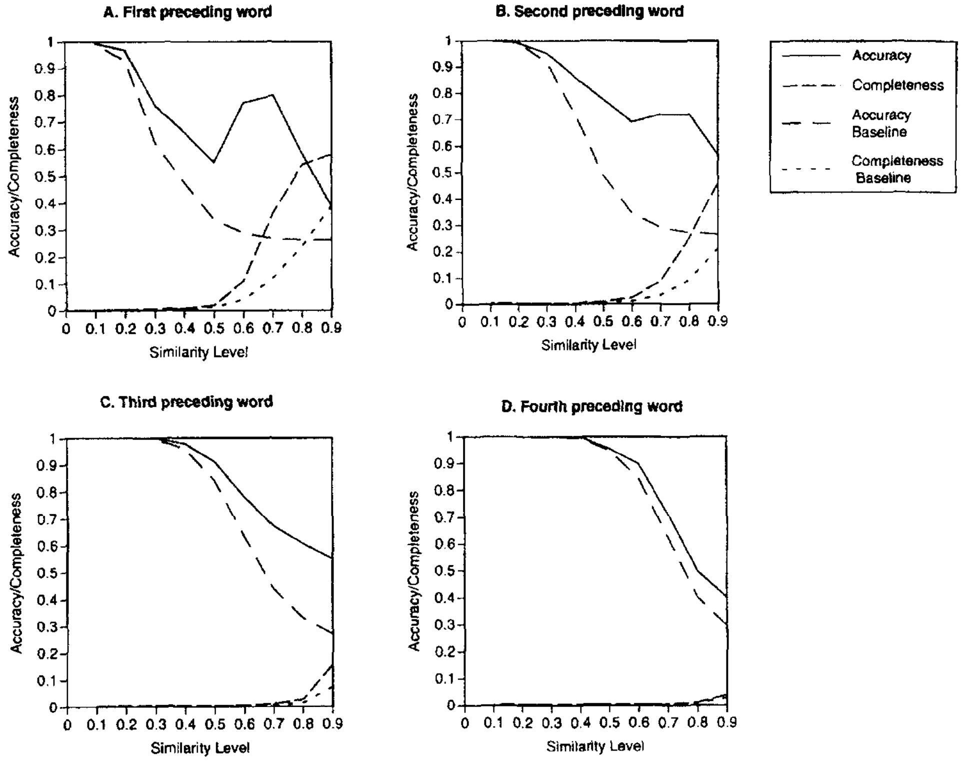

Figure 8. Accuracy and completeness when the first (A), second (B), third (C), and fourth (D) preceding words are used as context. The accuracy curves decrease as similarity increases, while the completeness curves increase. For both accuracy and completeness, the lower of the two lines is the random baseline, averaged over 10 random simulations. Standard deviations were too small to be shown.

for why a very wide context does not lead to improved results is that, for distant context positions, the syntactic relationships between target and context is highly variable, because the intervening material may have so many different possible syntactic structures.

Although preceding context was generally more informative than succeeding context, the best results were obtained by combining the two. Figure 9 shows accuracy and completeness when the first preceding and succeeding context positions, and the two preceding and two succeeding context positions were used as context. Here accuracy remains high at all levels of the dendrogram except the very highest. Increasing context to the four context positions either side of the target gave similar results for accuracy, but lower completeness, and comparing the one word either side and two words either side results shows that as context is broadened, completeness decreases. Overall, then, two context positions either side of the target word typically gives good results for both completeness and accuracy (.79 and .45, versus baselines of .27 and .15) at the 0.8 level of similarity. We therefore used this context in the further analyses reported below. The results of the analyses reported below can be regarded as an approximation of an upper bound on the utility of distributional

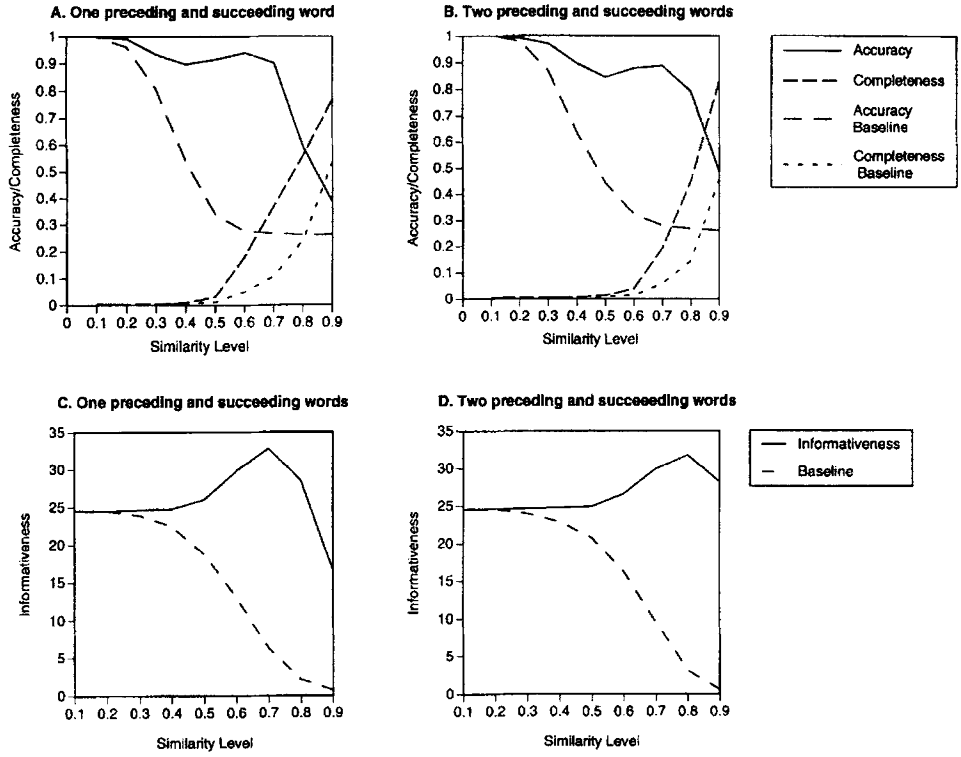

Figure 9. Accuracy and completeness when the first preceding and succeeding words (A), and the two preceding and succeeding words (B) are used as context. The accuracy curves decrease as similarity increases, while the completeness curves increase. For both accuracy and completeness, the lower of the two lines is the random baseline, averaged over 10 random simulations. Standard deviations were too small to be shown.

information. We do not expect that the qualitative pattern of results would differ substantially were some other highly local context employed.

For purposes of comparison, Figure 9 also shows the scores in terms of informativeness (calculated from Equation 3 above). Informativeness captures something of both the accuracy and completeness measures, and shows a inverted U-shaped profile, with maximum informativeness occurring at a intermediate level, close to the 0.8 level of similarity where the clustering is intuitively best. This measure reveals little difference between the clustering obtained from using two context positions to either side of the target and one context position to either side of the target.

Experiment 1 shows that highly local contexts are the most informative concerning syntactic category and that the amount of information they provide is considerable. These results have interesting general significance for the feasibility of distributional analysis. Although, as Pinker (1984) points out, there are an infinite number of possible distributional relationships between words, the very small number of highly local relationships, such as next word, preceding word, and preceding word but one, provide useful information about syntactic categories. Learners might be innately biased towards considering only these local contexts, whether as a result of limited processing abilities (e.g., see Elman, 1993) or as a result of a language-specific representational bias. From any viewpoint, information about highly local relationships between words could be picked up by a variety of plausible psychological mechanisms. Moreover, empirical evidence from a number of domains shows that both children and adults are highly sensitive to local dependencies in sequential material (Cleeremans, 1993; Saffran, Aslin & Newport, 1996). Thus, the fact that simple, highly local, relationships appear to be most informative about syntactic categories suggests that the hypothetical difficulties raised by Pinker do not appear to pose problems for distributional learning in practice.

Having examined the effect of using different context positions, we now consider how the number of target and context words affects the efficacy of distributional analysis.

Experiment 2: Varying the Number of Target and Context Words

The results above presuppose that the learner has a "vocabulary" of 1,000 target words, i.e., that the learner has to recognize 1,000 different target phonological strings. From a psychological standpoint, this raises two questions: What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child? To help answer these questions, we varied the number of words that were used as target and context items, keeping all other aspects of the analysis constant. In each analysis we used the most frequently occurring items, varying the number of these items that were used as target and context words. For reasons of space, we only sketch a general picture of the results here.

The effect of the number of target words can be characterized as an inverted U-shape. When the number of target words was very low, performance was quite poor, because the most frequent words tend to be closed-class words, for which (as we shall see below) this distributional learning method is less effective. As the number of target words grew to include many open-class words, and especially nouns and verbs, both accuracy and completeness increased. However beyond a certain number of target words accuracy and completeness tended to gradually decrease, as the additional words have relatively few occurrences and so the distributional statistics are less reliable. Using our entire 2.5 million word input corpus we observed a moderate decrease between 1,000 and 2,000 target words.

This is not to say that small numbers of target words cannot be successfully clustered—for example, Kiss's (1973) results, using only 31 target words show good qualitative results. However Kiss's 31 target words included many nouns and verbs. When we ran our standard analysis using the same 31 target words that Kiss used, accuracy and completeness (at the 0.8 level of similarity) were 0.50 and 0.78 (with random baselines of .27 and .44).

Varying the number of context words produced a similar inverted U-shape pattern. With a very small number of context words (e.g., 10) performance was relatively poor, as many target words did not cooccur with any of this small set. Increasing the number of context words to 50 resulted in a large gain in performance. Further increases beyond this point tended to trade increased accuracy for reduced completeness, and beyond 150 context words both accuracy and completeness degraded gracefully. With 500 context words the method was still providing some information, but the difference between the method and the random baseline was very small: with 1,000 target words accuracy and completeness were .40 and .44, with random baselines of .27 and .30.

In summary, then, the method works well even where there is a small vocabulary of target and context words, as long as the set of target words are largely content rather than function words. Although the child might not have access to 1,000 vocabulary items, if the child applies distributional analysis over its small productive vocabulary, this will work successfully, because this vocabulary consists almost entirely of content words. Moreover, prior to the vocabulary spurt, the child's syntax, and thus, presumably, knowledge of syntactic categories is extremely limited, and hence even modest amounts of distributional information may be sufficient to account for the child's knowledge. By the third year, the child's productive vocabulary will be approaching 1,000 items (e.g., Bates et al., 1994, found that the median productive vocabulary for 28 month olds was just under 600 words) and hence could in principle exploit the full power of the method.

It is also possible that, even when children's productive vocabularies are small, they may have a more extensive knowledge of the word forms in the language. It is possible that the child may be able to segment the speech signal into a large number of identifiable units, before understanding the meaning of the units (Jusczyk, 1997). Jusczyk and Aslin (1995) have shown that children who are exposed to novel words in isolation are able to recognize these in continuous speech, and moreover that children exposed to novel words in continuous speech can recognize them when presented in isolation. The child has no way of assigning a semantics to these novel words, but nonetheless appears to be sensitive to their occurrence, which suggests that they are represented as abstract word-forms. If this ability occurs more generally in language acquisition, such abstract units could be used as target or context items in a distributional analysis, because such analysis does not presuppose knowledge of meaning. This raises the possibility that a reasonably large-scale distributional analysis might occur even before the vocabulary spurt. Indeed, learning about syntactic classes at an early stage might be important even to a child with limited productive syntactic abilities, both for learning the syntax of its language, and in assigning meanings to word-forms (assuming a correlation between syntactic category and meaning).

Experiment 3: For Which Classes is Distributional Information of Value?

It is interesting to ask whether the distributional information about different syntactic classes obtained in our analyses reflects what is known about the order of children's acquisition of particular categories. In children, the major open classes, noun and verb, are believed to be acquired first, with the noun class preceding the verb class (Tomasello, 1992). Of course, this need not be a result of the use of distributional information. The use of semantic information would also predict this ordering of acquisition (as the referent of a noun is intuitively easier to discern than the referent of a verb, with both being easier than the "referent" of a function word). However, we can ask whether an advantage for open class items is consistent with the use of distributional information.

TABLE 2
The Major Categories from the Collins Cobuild Lexical Database, and Accuracy and Completeness Scores (and Random Baseline Scores) at the 0.8 Level of Similarity

		Observed		Baseline	
Class	n	Accuracy	Completeness	Accuracy	Completeness
noun	407	.90	.53	.43	.14
adjective	81	.38	.45	.09	.16
numeral	10	.09	.82	.02	.27
verb	239	.72	.24	.25	.14
article	3	.10	1.00	.01	.51
pronoun	52	.25	.24	.06	.14
adverb	60	.17	.18	.07	.16
preposition	21	.33	.53	.03	.16
conjunction	9	.06	.33	.02	.24
interjection	16	.18	.67	.02	.20
complex contraction	58	.55	.47	.07	.17
Overall	956	.72	.47	.27	.17

We therefore broke down the results of the standard analysis reported above by syntactic category (i.e., we used a context consisting of two words either side of the target, 150 context items, and 1,000 target items). Accuracy and completeness were calculated separately for the members of each syntactic category (according to the benchmark classification). Random baselines for each category were calculated similarly.

The observed accuracy and completeness, and the random baseline, when the dendrogram is cut at the 0.8 level of similarity, are shown in Table 2. Although this is a level of similarity that we chose by hand, this general pattern applies to the dendrogram as a whole. The most obvious feature of these results is that nouns emerge as the class for which distributional analysis provides the most information, which is consistent with developmental data. Performance for verbs is also impressive, although less good than nouns. Completeness for verbs is relatively low, reflecting the fact that in Figure 1 (the summary dendrogram for this level of similarity) verbs are broken into three clusters. At the higher similarity level of 0.9, completeness for verbs increases to .69 (baseline .46), with only a small decrease in accuracy (to .66, with baseline .25). For the remainder of the open class words, performance on adjectives is moderately good but adverb performance is relatively poor. Overall, better results are obtained for content words than for function words which is also consistent with developmental data.

The figures for numerals at this level of similarity are slightly misleading. As shown in Figure 6, numerals form a tight bunch with 100% accuracy at most levels of similarity, with accuracy and completeness of .99 and .82 at the 0.7 level of similarity (baselines of .03 and .21). Accuracy declines markedly at the 0.8 level of similarity only because the numerals then become grouped with the adjectives. The other slightly misleading figures are for the three word article class (consisting of *the*, *a*, and *an*). These are clustered with possessive pronouns, yielding perfect completeness (above the 0.5 level of similarity), but low accuracy. The figures for pronouns are probably an underestimate, as the score is reduced by the presence of many pronoun + auxiliary forms (e.g., *he'd*) in the same cluster, which are considered to be complex contractions by the benchmark.

In order to gain some intuition regarding why distributional information is more useful for content words than for function words, consider the kinds of contexts in which each will appear. Content words will tend to have one of a small number of function words as their context. Although content words are typically much less frequent, their context is relatively predictable. Function words, on the other hand, are much more frequent, but will tend to have content words as their context. Because there are many more content words, the context of function words will be relatively amorphous. As the measure of similarity exploits regularities in the distribution of contexts, those words with predictable contexts will be clustered together much more accurately.

Experiment 4: Corpus Size

Distributional analysis requires large amounts of data, and we therefore investigated whether such analysis can be effective using the amounts of language input typically available to the child. In fact, children are exposed to a remarkably large amount of language. Broen (1972) found that in free play with young children, mothers averaged 69.2 words per minute. A conservative minimum of one hour of free play per day would result in 1.5 million words of child-directed speech annually. Moreover, the child will, in addition, be exposed to a very much larger amount of non-child-directed speech. These figures are large in relation to the two and a half million words used in the analyses above. Although availability of language to the child is more than sufficient to support effective distributional learning, it is nonetheless interesting to ask how much input is required for the method to be effective.

In order to assess how much input was required to provide useful information, we ran analyses with 100,000 words, 500,000 words, 1 million words, and 2 million words of input. For all four analyses we used the two words either side of the target as context, and the most frequent 1,000 and 150 words as target and context words. In all analyses the method provided more information than the random baseline. However this advantage was very slight for the 100,000 words simulation (at the 0.8 level, accuracy and completeness were .60 and .01, against baselines of .47 and .01). With 500,000 of input the advantage of the distributional analysis was more marked (.43 and .26 versus baselines of .26 and .17), but it was with 1,000,000 words of input that the method really took off, with accuracy and completeness of .72 and .42, (against baselines of .27 and .15). From here performance increased gradually with increasing amounts of input, so that the results we report for the full analysis are not a genuine upper bound: Given more input, it seems likely that some small further increase in performance could be expected.

It is interesting to note that distributional learning of syntactic categories can nonetheless be useful with much less input. If the target items are open class, it is possible to divide items into nouns and verbs even with as few as 15,000 words of child-directed speech (Kiss, 1973). This indicates that distributional learning may be useful at least to some extent, even if the child disregards the overwhelming majority of the speech to which they are exposed.

Experiment 5: Utterance Boundaries

An unrealistic feature of our distributional analyses reported above is that they concatenate language from different speakers into a single undifferentiated speech stream. But the child knows when one utterance ends, and another begins. This knowledge of utterance boundaries might potentially provide additional information about syntactic categories. For example, utterances typically end with content, rather than function, words, and single word utterances tend to be nouns, especially names, interjections or more rarely other content words. Moreover, distributions that are measured across utterances (particularly when there is a change of speaker) would be likely to contain a great deal of noise, because syntactic constraints do not apply across utterances. We therefore repeated the standard analysis above, but added utterance boundaries information.

We incorporated utterance boundaries in two ways. First, we simply did not record context items which were not in the same utterance as the target. This should reduce the amount of noise in the analysis. This way of including utterance boundary information may be criticized, however, because it does not exploit potential information from single word utterances—it records nothing from those utterances. Hence, we also used a second method, where utterances boundaries are treated as an additional lexical item—that is, they are explicitly present in the corpus, and the utterance boundary marker is also used as a context item.

Figure 10 shows the informativeness of classifications from our standard analysis, when context was not measured across utterance boundaries, and when utterance boundaries were explicitly marked. We used the informativeness measure (described above) in order to make the comparison between classifications clearer. At the 0.8 level of similarity, our

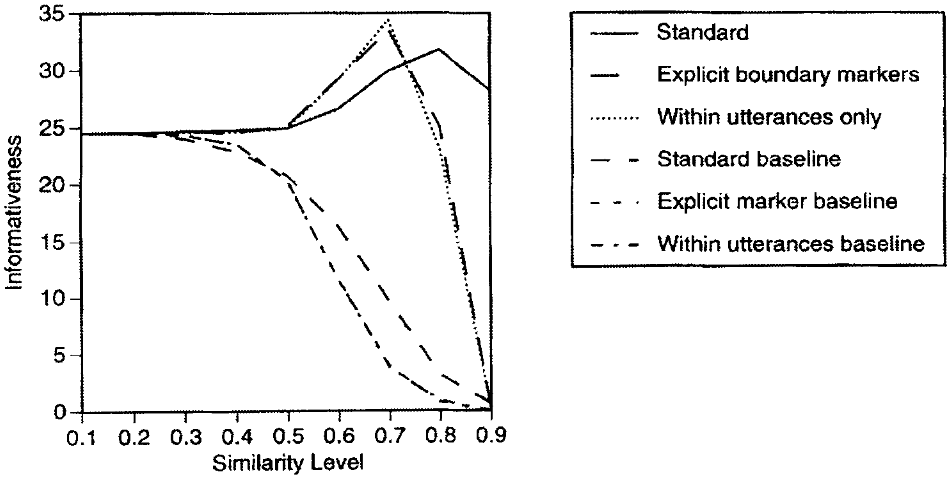

Figure 10. The informativeness of the standard analysis, the same analysis without measuring context across utterance boundaries, and analysis where utterance boundaries were explicitly marked. Random baselines (as discussed in the text) are also shown for all three analyses.

standard analysis provides more information than the analyses taking utterance boundaries into account. However, when the 0.7 level of similarity the advantage of the latter analyses is clear. As well as improving the informativeness of the classification, the analyses that take account of utterance boundaries shift the best level of the dendrogram (where it best captures the benchmark classification) to a lower level of similarity. It appears that the information recorded across utterance boundaries effectively acts as noise. Removing this information improves classification, but marking utterance boundaries provides very little extra benefit.

The results above suggest that although the child could (and presumably does) use utterance boundary information to constrain distributional analysis, this is not critical to the effectiveness of the distributional analysis we propose.

Experiment 6: Frequency Versus Occurrence

The analyses that we have outlined assume that the child is sensitive to the (rank order of) frequencies with which a target word is paired with different context items. Although it is plausible that children are sensitive to frequency information, it is interesting to wonder whether distributional methods where information on occurrence, but not *frequency* of occurrence, is recorded. We therefore repeated the standard analysis, but replaced all nonzero frequency counts between target and context items with the number 1, indicating that the target and context items were observed together.

A slight complication to this analysis is that the Spearman rank correlation coefficient is not a good similarity metric between binary vectors, where there are only two possible ranks. Therefore we used the overlap between vectors, known as the city-block metric, as a measure of similarity.

Figure 11 shows the results of the standard analysis (using frequency information and the Spearman correlation), of an analysis using frequency information and the city-block metric, and of an analysis using only occurrence and the city-block metric. The results clearly show that using the city-block measure (instead of the rank correlation) leads to a major decrement in performance. Nevertheless, the method still provides a considerable amount of information about the syntactic categories of the target words. When frequency information is excluded altogether, however, the amount of information provided by the method is very small, although still slightly greater than the random baseline.

Thus it appears that cooccurrence information could still be used to constrain words' syntactic categories in the absence of frequency information, but that this distributional method works very much better when frequency information is included.

Experiment 7: Removing Function Words

Early child speech consists largely of content rather than function words (Bloom, 1970). This suggests the possibility that children might pay much more attention to these words in comprehension. If the child only attends to content words, then the language input available to the mechanism for acquiring syntactic categories may consist of a stream of content words, with the function words effectively "edited out." To

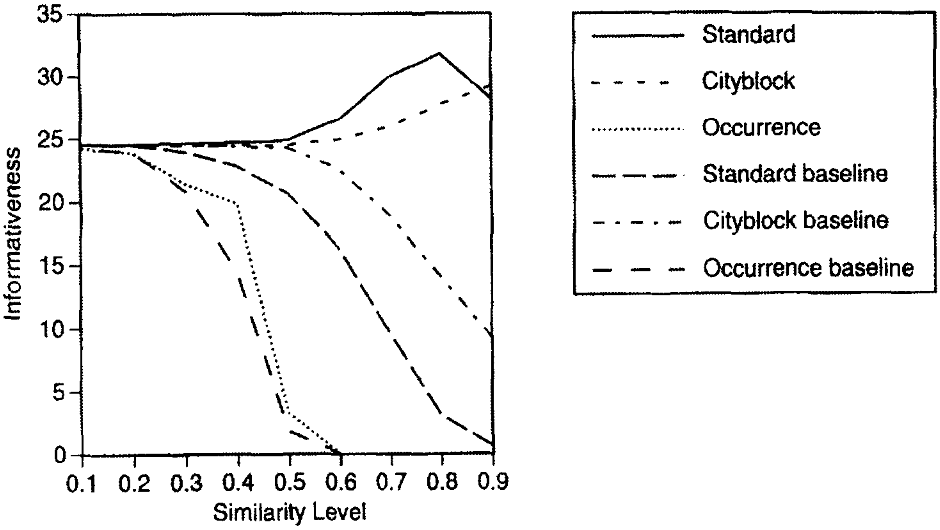

Figure 11. The informativeness of the standard analysis, the standard analysis using the cityblock metric for similarity, and a third analysis using the cityblock metric, and occurrence (as opposed to frequency of occurrence) to measure context. Random baselines (as discussed in the text) are also shown for all three analyses.

explore whether distributional learning would be successful in these circumstances, we stripped out the function words from our corpus, but otherwise ran the analyses as before.

Figure 12 shows the informativeness of the resulting dendrogram with respect to the benchmark, compared against the informativeness of the standard analysis. Although removing function words does have a considerable impact on the amount of information provided by the method, the analysis still provides a considerable amount of useful information.

Experiment 8: Does Information About One Category Help the Acquisition of the Others?

The analyses so far have embodied the assumption that the syntactic categories of all lexical items are derived simultaneously, using only distributional information. But, as we noted at the outset, the child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis. In particular, it is interesting to ask to what extent "hints" from other sources of information about a certain class of items might assist distributional analysis for the other classes. For example, semantic information may help the child identify the class of nouns, or the class of verbs, before the other syntactic classes have been identified. Alternatively, frequency information or the lack of an identifiable referent, might allow the child to group together function words (even though the child may be unable to classify between function words, or to understand their syntactic or

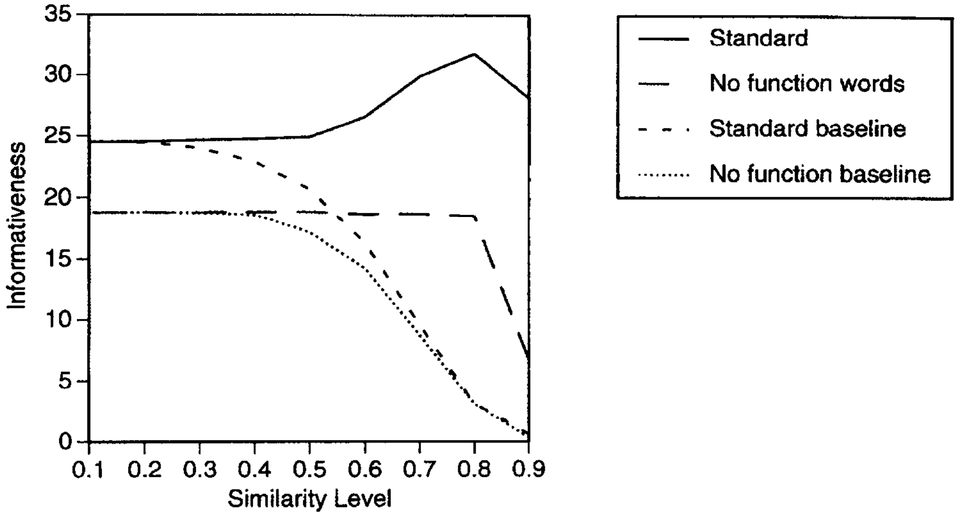

Figure 12. The informativeness of the standard analysis, and when function words are completely excluded from the analysis. Random baselines (as discussed in the text) are also shown.

semantic roles). To what extent would knowing these classes aid in learning about the other syntactic categories.[11]

One way of using information about a particular class in our distributional analysis is by replacing all words of a particular category with a symbol representing that category. For example, all nouns might be replaced with the category label NOUN. This has the advantage that contexts concerning different nouns can be identified as having the same syntactic significance. But it also has the potential disadvantage that information about differences between nouns (e.g., singular versus plural, count vs. mass nouns, and so on) is lost. We repeated our standard analysis with three variants: all nouns replaced by the symbol NOUN, all verbs replaced by the symbol VERB, and all function words (articles, pronouns, prepositions, and conjunctions) replaced by the symbol FUNCTION. The results of this analysis are shown in Figure 13.

Perhaps surprisingly, all of these analyses reveal a slight decrement in performance with respect to the amount of syntactic information concerning the members of other syntactic classes. For example, when all nouns are grouped together, the remaining words are less accurately discriminated, and vice versa. Moreover, content words are slightly less well discriminated when function words are classified together. These results suggest that it may not be appropriate to integrate information from other sources into the distributional analysis by collapsing sets of lexical items (and their frequency counts) into discrete categories. That is, frequency information about individual lexical items in all classes may be important in exploiting distributional information. An interesting topic for future research is to explore other ways in which non-distributional sources of information might be integrated with a distributional analysis.

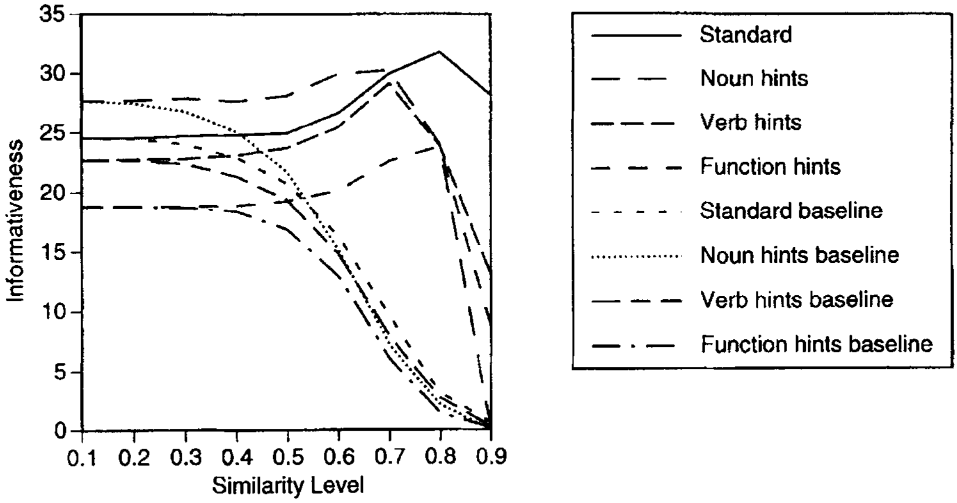

Figure 13. Informativeness and baseline values for the standard analysis, when all nouns are replaced by a single symbol ("noun hints"), when all verbs are replaced by a single symbol, and when all function words (articles, pronouns, prepositions, and conjunctions) are replaced by a single symbol.

Experiment 9: Is Learning Easier with Child-Directed Input?

The analyses reported in this paper have been conducted using a corpus of speech to which children were exposed. Much research on child-directed speech has stressed that language used to children is different in terms of both vocabulary and syntactic complexity from normal adult-adult speech. Many researchers have suggested that "motherese" represents an adaptation of the speech input to facilitate learning. This suggests the possibility that one function of motherese may be to enhance the acquisition of syntactic categories. We therefore investigated whether the present distributional learning analyses are sensitive to the difference between adult speech to children, and adult-adult speech.

We compared our standard analysis on the CHILDES corpus with a similar analysis on a corpus of conversational adult-adult speech taken from the British National Corpus (BNC). This material was recorded by the speakers in a variety of informal settings in the United Kingdom. Although some of the materials may have involved speech to children, the vast majority was comprised of adult-adult conversation. Like the CHILDES corpus, the BNC is a large noisy corpus, with many speakers, dialects, constructions, topics, and vocabulary items. We extracted speech from this corpus, selecting files (each containing a transcription of a single session) at random, until we had a subset of the corpus which was equal in size to the CHILDES corpus (2.57 million words). This corpus was preprocessed in exactly the same manner as the CHILDES corpus (all punctuation and case information was removed, etc.). We then performed our standard distributional analysis on the corpus, using the most frequent 1,000 words as targets, and the most frequent 150 words as context, using two context positions to either side of the target word. The benchmark categorisation for the BNC target words was derived from the CELEX database, in exactly the

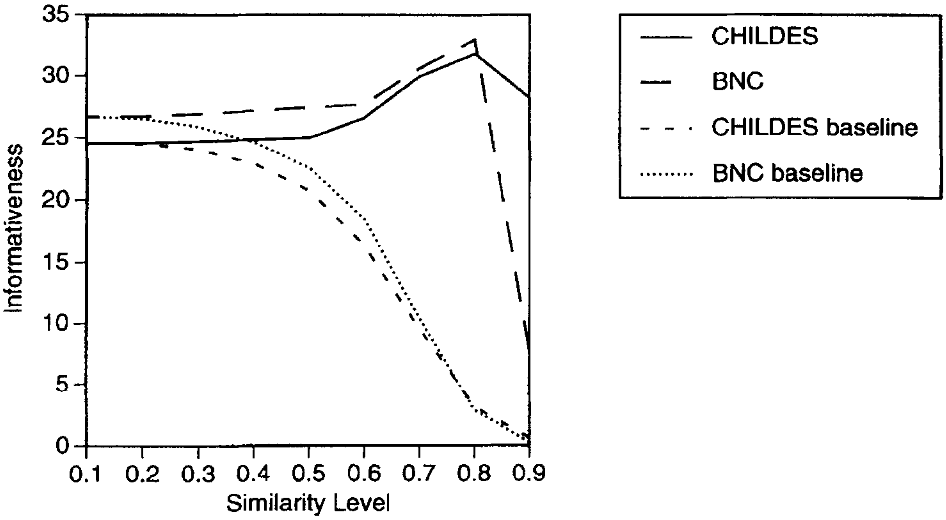

Figure 14. Informativeness and baseline values for the standard analysis, on the CHILDES corpus, and on materials from the British National Corpus (BNC).

same manner as for the CHILDES corpus. The results of the BNC and standard CHILDES analysis are shown in Figure 14.

Clearly there is very little difference between the two analyses, with, if anything, a slight advantage for the adult-adult speech analysis. This is quite surprising, given the relatively simplified nature of parental speech to children, which should in principle provide more reliable statistics for the target words. One possible explanation for this effect is that the adult language in the CHILDES corpus was dominated by adult-adult speech (even though this speech was recorded in the presence of children, which is not generally true of the BNC material). Regardless, this finding suggests that the distributional mechanism is not dependent on motherese, and is consistent with evidence that children whose carers do not use motherese do not acquire language any less quickly (e.g., Newport, Gleitman, & Gleitman, 1977).

III. DISCUSSION

We have proposed a model of how children use distributional information to constrain the acquisition of syntactic categories. This model uses highly local distributional information, concerning the immediately preceding and succeeding items surrounding the target word (Experiment 1), is consistent with what is known about early vocabulary development in general (Experiment 2) and is most effective for learning nouns, and then verbs, and least effective for function words, mirroring children's syntactic development (Experiment 3). The method learns using the input corpora of the order of magnitude received by the child (Experiment 4), works whether or not utterance boundaries are explicitly marked (Experiment 5), and crucially gains massive benefits from knowledge of the frequency of occurrence (rather than the mere occurrence) of distributional relationships between pairs of words (Experiment 6). The method still works, although slightly less well, when function words are removed from the input (Experiment 7), and when sets of words are replaced in the corpus by a syntactic category label (Experiment 8), and is as effective with adult-adult speech as with the adult speech taken from the CHILDES corpus, containing a significant proportion of child-directed speech (Experiment 9). The success of this distributional model of syntactic category acquisition suggests that distributional information may make an important contribution to early language development. We now consider possible extensions of this work.

We have explored the feasibility of a particular distributional method on a corpus appropriate to the acquisition of syntactic categories for English. How widely does this particular method apply? It seems likely that it will be most successful for languages which have strong word order constraints, since it uses sequential order information. Nonetheless, even in many languages (such as Italian and Russian) in which word order is relatively free in principle, there are somewhat stereotypical, though not obligatory, word order patterns, which might suffice for some degree of success. An obstacle to a thorough cross-linguistic study is that large machine-readable corpora are readily available for only a small number of languages. One important line of future research is to apply this and similar methods to corpora of other languages as they become available.

It seems likely that other distributional measures or analyses may provide valuable clues to syntactic classes, and that these might be used either in conjunction with, or instead of, the particular approach we outline here. For example, as discussed above, Maratsos (1988) has suggested that a distributional analysis based on morphological cues might be a valuable source of such information. In, for example, languages which rely heavily on case-marking (e.g., Turkish), this may be a better source of information than the kind of distributional analysis we have developed, which depends on word order constraints. To apply such a distributional analysis requires that morphemes can be identified, and this may also be possible by using distributional methods (e.g., Brent, 1993). It is also possible that the importance of distributional information, as a whole, varies greatly from one language to another. Although all languages may be similar at a deep level (e.g., Chomsky, 1980), they appear very different at a superficial level. Hence we would expect that methods and sources, both distributional and otherwise, used in finding linguistic information from this superficial level will likewise be highly variable across languages.

The general question of how distributional information can be integrated with other sources is an important one. In particular, phonological information is an obvious candidate for both isolated and combined study, given the evidence concerning its value as a cue to syntax (Kelly, 1992), and the relative simplicity of applying similar methods to phonologically transcribed corpora such as the Lund corpus (Svartvik & Quirk, 1980; see Shillcock, Hicks, Cairns, Levy & Chater, in press). Again, the relative paucity of such corpora (especially of languages other than English) is a current obstacle to this line of enquiry. But it seems likely that constraints from a range of sources should make the learning problem significantly easier (see Christiansen, Allen, Seidenberg, in press, for related discussion).

The fine-grained structure of the results above shows considerable semantic influence. It is possible that distributional information may have some influence on learning word meanings. Indeed, Gleitman (1994) argues that syntactic information (which, we have argued, could potentially be acquired by distributional methods) might be important in the acquisition of verb meanings. Within computational linguistics, various researchers (Schutze, 1993; Tishby & Gorin, 1994) have found some degree of semantic relatedness between words using simple distributional methods. More psychologically oriented work has also been conducted by Lund and Burgess (1996).

One limitation of the current model, is that it takes no account of the fact that many words are syntactically ambiguous. It is interesting to consider how the model might be extended to allow this possibility. It is possible that given an initial assignment of categories, constrained by the distributional analysis above, other categories can be learned by observing the variety of distributions in which a word occurs. For example, consider a word which is sometimes a noun, but very much more frequently a verb, and hence is clustered with other verbs. When this word is observed in a stereotypical noun context, it may be possible to infer that it can also function as a noun. One piece of evidence that this may be possible is that comparing context vectors based on a single occurrence with those averaged over many contexts may give significantly above chance assignment of syntactic category (Redington, Chater & Finch, 1993). The critical question regarding the viability of this approach is whether or not genuine secondary readings can be obtained without contamination from other spurious readings. Additionally, of course, many other cues or methods might be used to identify a word's multiple syntactic categories.[12]

IV. CONCLUSIONS

We have proposed a model of how children may use distributional information in acquiring syntactic categories. Independent of the details of our specific proposal, our results show that distributional information is a potentially powerful cue for learning syntactic categories. No similarly successful demonstration of the computational value of any other source of constraint on syntactic categories, whether phonological, prosodic or semantic, has been provided.

The use of distributional methods is often associated with empiricist approaches to language acquisition. As should be clear, our stance is neutral regarding this wider debate. We believe that distributional analysis may be a source of useful information in acquiring many aspects of language (see Redington & Chater, 1997, in press), but this by no means implies that many other sources of information, including innate constraints, are not crucially important. The use of distributional information is consistent with any point on the nativist-empiricist continuum. Language acquisition poses difficult problems both for the child and for the researcher. By focusing on simple aspects of language, and simple approaches to how they may be learned, as we have done here, it may be that some concrete progress may be possible. How far such simple methods may succeed in finding other aspects of linguistic structure is a matter for further empirical investigation.

This work is an example of a general approach, which focuses on quantitatively assessing the potential contributions of information sources concerning aspects of linguistic structure. As well as providing feasibility proofs concerning the utility of different sources of information, this research promises to form the basis for computationally explicit models of specific aspects of language acquisition.

Acknowledgements: This research was supported by SERC Studentship 91425111, and ESRC Research Studentship R00429234268, and Research Grant SPG9029590 from the Joint Councils Initiative in Cognitive Science/HCI. We would like to thank the staff of the CHILDES project, and Claire Warwick for help with the British National Corpus, and the Center for Research in Language, University of California, San Diego, where some of this research has been conducted, Paul Cairns for supplying the canonical categories from the Collins Cobuild database, Matt Crocker for guidance and advice, and Mark Seidenberg, Michael Brent, and two anonymous reviewers for their detailed comments on an earlier draft of this paper.

NOTES

[1] In the context of syntactic category acquisition, this usage is standard in the literature. In other contexts, distributional information may be interpreted more broadly, to include, for example, the relation between a word and its phonological constituents, or its extra-linguistic contexts. This wider usage is inappropriate here because it conflates the different sources of information that may be useful in learning syntactic categories, from phonological to semantic constraints.

[2] Pinker (1984) sees distributional analysis as potentially useful only when constrained by semantic information.

[3] Indeed, it is quite possible that the interaction of information sources may be crucial, if individual sources are relatively weak when considered alone (see e.g., Christiansen, Allen & Seidenberg, in press).

[4] We thank two anonymous reviewers for pointing out this objection.

[5] See Redington & Chater, in press, for further discussion of these and further objections to the role of distributional methods in language acquisition in general.

[6] For discussion of the close relationship of neural network and statistical methods see, for example, Chater (1995).

[7] Specifically, the hidden unit activations associated with each occurrence of each item are averaged together. A measure of similarity between each pair of items can then be derived by the Euclidean distance between the average hidden unit values for each lexical item. The resulting matrix of similarities can then be cluster-analysed (Elman, 1990). Similar results have also been obtained without averaging hidden unit activations across contexts (Elman, 1990).

[8] It is not clear whether a similar analysis applies when SRNs are trained on more complex grammars, involving, for example, aspects of recursion (Christiansen & Chater, 1997; Elman, 1991; Weckerly & Elman, 1992).

[9] We stress that, of course, full natural language syntax is immeasurably richer than sequential material which has only bigram structure. Bigram statistics function here as a clue to aspects of syntactic structure, not as a model of the language.

[10] Given another source of information about syntactic classes (e.g., phonological or semantic cues) the child could use the detection of this "humped" pattern in determining at which level of similarity best corresponds to syntactic categories.

[11] This analysis was suggested to us by Jenny Saffran, in relation to similar work by Mintz, Newport and Bever (1995).

[12] Recent work using a very different distributional method by Cartwright and Brent (1997) aims to address this question.

REFERENCES

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21,* 85-124.

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

Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. *Proceedings of the 15th Meeting of the Cognitive Science Society* (pp. 28-36) Hillsdale, NJ: LEA.

Brent, M. R. & Cartwright, T. A. (1997). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition,* 63, 121-170.

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

Brill, E., Magerman, D., Marcus, M. & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. *DARPA Speech and Natural Language Workshop.* Hidden Valley, PA: Morgan Kaufmann.

Broen, P. A. (1972). *The Verbal Environment of the Language Learning Child.* Washington, DC: Asha.

Bruner, J. (1975). The ontogenesis of speech acts. *Journal of Child Language, 2,* 1-19.

Cairns, P., Shillcock, R. C., Chater, N. & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns (Eds.), *Connectionist Models of Memory and Language,* (pp. 289-310). London: UCL Press.

Cartwright. T. A. & Brent, M. R. (1997). Early acquisition of syntactic categories: A formal model. *Cognition, 63,* 121-170.

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

Charniak, E. (1993). *Statistical Language Learning.* Cambridge, MA: MIT Press.

Chater, N. (1995). Neural networks: The new statistical models of mind. In J. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns (Eds.), *Connectionist Models of Memory and Language,* (pp. 207-227). London: UCL Press.

Chater, N. & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. *Proceedings of the 14th Meeting of the Cognitive Science Society* (pp. 402-407) Hillsdale, NJ: LEA.

Chater, N. & Conkey, P. (1993). Sequence processing with recurrent neural networks. In M. Oaksford & G. D. A. Brown (Eds.), *Neurodynamics and Psychology.* London: Academic Press.

Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language, 35,* 26-58.

Chomsky, N. (1964). *Current Issues in Linguistic Theory.* The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Boston, MA: MIT Press.

Chomsky, N. (1980). *Rules and Representations.* Cambridge, MA: MIT press.

Christiansen, M. H. (in preparation). *Natural language recursion and recurrent neural networks.*

Christiansen, M. H. & Chater, N. (1997). *Toward a connectionist model of recursion in human linguistic performance.* Manuscript submitted for publication.

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

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. *Proceeding of the Second Conference on Applied Natural Language* (pp. 136-143) Austin, TX.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning.* Boston, MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation, 1,* 372-381.

Conkey, P. (1991). *Sequence prediction using recurrent neural networks.* MSc Thesis, Department of Artificial Intelligence, University of Edinburgh.

Davis, S., Morris, J. & Kelly, M. H. (1992). *The causes of duration differences between English nouns and verbs.* Unpublished manuscript.

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

Elman, J. L. (1991). Distributed representations, simple recurrent neural networks, and grammatical structure. *Machine Learning, 7,* 195-225.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48,* 71-99.

Fernald, A. (1994). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In P. Bloom (Ed.), *Language Acquisition: Core Readings*. Cambridge, MA: MIT Press.

Finch, S. (1993). Finding Structure in Language. Ph.D. Thesis, Centre for Cognitive Science, University of Edinburgh.

Finch, S. P. & Chater, N. (1991). A hybrid approach to the automatic learning of linguistic categories. *AISB Quarterly, 78,* 16-24.

Finch, S. P. & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the 14th Annual Conference of the Cognitive Science Society of America* (pp. 820-825) Bloomington, IN.

Finch, S. P. & Chater, N. (1993). Learning syntactic categories: A statistical approach. In M. Oaksford, & G. D. A. Brown, (Eds.), *Neurodynamics and Psychology*. London: Academic Press.

Finch, S. P. & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society.*

Fries, C. C. (1952). *The Structure of English. London: Longmans.* Garside, R., Leech, G. & Sampson, G. (1987). The Computational Analysis of English—A Corpus Based Approach. London: Longman.

Gleitman, L. (1994). The structural sources of verb meanings. In P. Bloom (Ed.), *Language Acquisition: Core Readings*. Cambridge, MA: MIT Press.

Gleitman, L. R., Gleitman, H., Landau, B. & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge Survey, Vol. 3.* Cambridge: Cambridge University Press.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. L. Baker, and J. McCarthy (Eds.), *The Logical Problem of Language Acquisition.* Cambridge, Mass: MIT Press.

Harris, Z. S. (1954). Distributional structure. *Word, 10,* 140-162.

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31,* 190-222.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks.* New York: Wiley.

Hirsh-Pasek, K., Kemler-Nelson, D. G., Jusczyk, P. K., Wright, K. & Druss, B. (1987). Clauses are perceptual units for prelinguistic infants. *Cognition, 26,* 269-286.

Ingram, D. (1989). *First Language Acquisition: Method, Description & Explanation.* Cambridge: Cambridge University Press.

Jusczyk, P. K. (1993). Discovering sound patterns in the native language. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* (pp 49-60) Hillsdale, NJ: LEA.

Jusczyk, P. K., Cutler, A. & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64,* 675-687.

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

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

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

Kuczaj, S. A. (1982). On the nature of syntactic development. In S. A. Kuczaj (Ed.), *Language Development, Volume 1: Syntax and Semantics.* Hillsdale, NJ: LEA.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language, 6,* 225-242.

Jusczyk, P. W. (1997) *The discovery of spoken language.* Cambridge, MA: MIT Press.

Jusczyk, P. W. & Aslin, R. N. (1995). Infants' detection of sound patterns in fluent speech. *Cognitive Psychology, 29,* 1-23.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Levy, Y. (1983). It's frogs all the way down. *Cognition, 15,* 75-93.

Levy, Y, & Schlesinger, I. M. (1988). The child's early categories: Approaches to language acquisition theory. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine, (Eds.), *Categories and Processes in Language Acquisition,* Hillsdale, NJ: LEA.

Liberman, M. & Prince, A. S. (1977). On the stress and linguistic rhythm. *Linguistic Inquiry,* 8, 249-336.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-ocurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203-208.

MacWhinney, B. (1989). *The CHILDES Project: Computational Tools for Analyzing Talk.* Hillsdale, NJ: LEA.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12,* 271-295.

Maratsos, M. (1979). How to get from words to sentences. In D. Aaronson & R. Rieber (Eds.), *Perspectives in Psycholinguistics.* Hillsdale, NJ: LEA.

Maratsos, M. (1988). The acquisition of formal word classes. In Y. Levy, I. M. Schlesinger & M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

Maratsos, M. & Chalkley, M. (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.

Marcus, M. (1991). The automatic acquisition of linguistic structure from large corpora. In D. Powers (Ed.), *Proceedings of the 1991 Spring Symposium on the Machine Learning of Natural Language and Ontology,* Stanford, CA.

Mintz, T. H., Newport, E. L., & Bever, T. G. (1995). Distributional regularities of grammatical categories in speech to infants. *Proceedings of the 25th Annual Meeting of the North Eastern Linguistics Society.* Amherst, MA.

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

Nelson, K. (1977). Facilitating children's syntax acquisition. *Developmental Psychology, 13,* 101-107.

Newport, E. L., Gleitman, H. R. & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow and C. A. Ferguson (Eds.) *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. (1975). *Statistical Package for the Social Sciences.* 2nd edition. New York: McGraw Hill.

Ninio, A. & Snow, C. E. (1988). Language acquisition through language use: The functional sources of children's early utterances. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

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

Popova, M. I. (1973). Grammatical elements of language in the speech of pre-school children. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of Child Language Development.* New York: Holt, Rinehart, & Winston.

Radford, A. (1988). *Transformational grammar,* 2nd Edition. Cambridge: Cambridge University Press.

Redington, M. & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences, 1,* 273-281.

Redington, M. & Chater, N. (in press). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes.*

Redington, F. M., Chater, N. & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* (pp 848-853) Hillsdale, NJ: LEA.

Ritter, H. & Kohonen, T. (1989). Self-organizing semantical maps. *Biological Cybernetics,* 62, 241-254.

Ritter, H. & Kohonen, T. (1990). Learning "semantotopic maps" from context. *Proceedings of the International Joint Conference on Neural Networks, Vol. 1,* 23-26.

Rosenfeld, A., Huang, H. K. & Schneider, V. B. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information Theory, 15,* 672-681.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 376-380) Mawah, NJ: Lawrence Erlbaum Associates.

Schlesinger, I. M. (1981). Semantic assimilation in the acquisition of relational categories. In W. Deutsch (Ed.), *The Child's Construction of Language.* New York: Academic Press.

Schlesinger, I. M. (1988). The origin of relational categories. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

Scholtes, J. C. (1991a). Kohonen's self-organising map applied towards natural language processing. *Proceedings of the CUNY 1991 Conference on Human Sentence Processing.*

Scholtes, J. C. (1991b). Using extended feature maps in a language acquisition model. *Proceedings of the 2nd Australian Conference on Neural Networks.*

Schutze, H., (1993). Word Space. In S. J. Hanson, J. D. Cowan, & C. L. Giles (eds.), *Advances in Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann.

Shillcock, R., Hicks, J., Cairns, P., Levy, J. & Chater, N. (in press). A statistical analysis of an idealised phonological transcription of the London-Lund corpus. *Computer Speech and Language.*

Shillcock, R. C., Lindsey, G., Levy, J. & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. *Proceedings of the 14th Annual Meeting of the Cognitive Science Society* (pp 408-414) Hillsdale, NJ: LEA.

Snow, C. E. (1972). Mother's speech to children learning language. *Child Development, 43,* 549-565.

Snow, C. E. (1988). The last word: Questions about the emergence of words. In M. Smith and J. Locke (Eds.), *The Emergent Lexicon.* New York: Academic Press.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco: W. H. Freeman.

Sorenson, J. M., Cooper, W. E. & Paccia, J. M. (1978). Speech timing of grammatical categories. *Cognition, 6,* 135-153.

Svartvik, J. & Quirk, R. (1980). *A corpus of english conversation.* Lund: LiberLaromedel Lund.

Taylor, J. R. (1989). *Linguistic categorization: Prototypes in linguistic theory.* Oxford: Clarendon Press.

Tishby, N. & Gorin, A. (1994). Algebraic learning of statistical associations. *Computer Speech & Language, 8,* 51-78.

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

Tucker, G. R., Lambert, W. E., Rigault, A. & Segalowitz, N. (1968). A psychological investigation of French speakers' skill with grammatical gender. *Journal of Verbal Learning and Verbal Behavior, 7,* 312-316.

Weckerly, J. & Elman, J. (1992). A PDP approach to processing center-embedded sentences. *Proceedings of the 14th Meeting of the Cognitive Science Society* (pp. 414-419) Hillsdale, NJ: LEA.

Wolff, J. G. (1976). Frequency, conceptual structure and pattern recognition. *British Journal of Psychology, 67,* 377-390.

Wolff, J. G. (1977). The discovery of segmentation in natural language. *British Journal of Psychology, 68,* 97-106.

Wolff, J. G. (1988). Learning syntax through optimisation and distributional analysis. In Y. Levy, I. M. Schlesinger and M. D. S. Braine (Eds.), *Categories and Processes in Language Acquisition.* Hillsdale, NJ: LEA.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston, MA: Houghton Mifflin.