Connectionist Models of Memory and Language

Edited by

Joseph P. Levy, Dimitrios Bairaktaris,
John A. Bullinaria, Paul Cairns

UCL
PRESS

CHAPTER 12

Acquiring syntactic information from distributional statistics

Steve Finch, Nick Chater, Martin Redington

Introduction

Acquiring syntax appears to face the language learner with a "bootstrapping" problem. Acquiring syntactic rules presupposes that the syntactic categories in terms of which those rules are formulated have already been acquired; but syntactic categories only have meaning in virtue of the syntactic rules in which they figure. Learning syntactic categories and syntactic rules appear to be mutually interdependent. Consequently, the learner appears to be faced with what seems an impossible task: searching the entire space of category/rule combinations simultaneously.

Even if, as many theorists assume, the learner is equipped with a rich innate knowledge of grammatical rules and abstract grammatical categories, the problem of mapping lexical items onto such categories still remains. Indeed, only once the learner has learned to categorize the speech stream in terms of relatively complex categories will it be possible to bring any innate grammatical information to bear on the learning process. For example, the learner cannot assess word order constraints in the target language, or whether or not some linguistic feature such as "pro-drop" is allowed, until the linguistic input is represented in terms of distinct words each labelled with (at least an approximation to) its syntactic category. Therefore it seems that whether or not learners possess an innate store of grammatical information, the initial stages of language acquisition must be driven by linguistic input.

In considering the potential contribution of sources of information in the linguistic input, two questions may be asked: Can the putative source of information contribute in principle, and how could the relevant information be obtained? And is there empirical evidence to suggest that infants utilize the source of information? In this chapter, we are concerned almost exclusively with the first question, and hence our focus will be on the possible utility of different kinds of other cues, which need not be useful on theoretical grounds, may in practice co-occur 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 both extremely labour-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 to 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 labour-intensive to notate. There are currently no large (millions of words) corpora of conversation with detailed prosodic markings. In 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.

Distributional methods can often be readily applied to language internal information, since word level corpora exist. In particular, unlike semantic and prosodic approaches, distributional analysis can be conducted over texts, represented purely as sequences of distinct words, and these are (at least for English) in almost unlimited supply. Also, reasonably large corpora of transcribed speech, such as the London–Lund corpus (Svartvik & Quirk 1980) and the CHILDES database (MacWhinney & Snow 1985) are also available. These are at least large enough to provide some validation of the performance of distributional methods which are primarily developed using text corpora.

In this chapter, we shall describe and illustrate the performance of a simple distributional technique. This has been applied to learning syntactic categories of lexical items and two- and three-word phrases (Finch & Chater 1992, Finch & Chater 1993), and here we report an extension of this approach to more complex phrasal structure.

While we focus on the distributional approach, it seems entirely likely that all of the types of information source (including semantic and prosodic sources) may be (perhaps highly) informative about syntactic structure and that, if so, the child may draw on them. It is simply that, for the reasons outlined above, such questions are very hard to investigate. Thus, we restrict the discussion here to quantitatively considering distributional methods on purely methodological grounds. For the same reasons, it is difficult to assess the potential importance of interactions between these sources and distributional information. Although we believe that such interactions may be very important, here we shall consider the potential role of distributional information in isolation. Notwithstanding, it is quite possible that the interaction of information sources is so important that any individual source is relatively weak when considered alone.

Objections to the feasibility of distributional methods

A number of arguments have been put forward which appear to undermine the feasibility of distributional methods (in isolation[1]) in category acquisition, and in syntax acquisition in general. Pinker (1984, 1987, 1989) makes the most cogent and influential case against distributional methods. He argues that these criticisms will apply to all distributional methods, illustrating his arguments by considering the work of Maratsos & Chalkley (1981). Pinker suggests that distributional methods have two fundamental problems. His *learnability* argument aims to establish that distributional methods are inadequate in principle, and his *efficiency* argument aims to show that they are unworkable in practice:

(a) *Learnability*. Since distributional methods work solely by examining observed utterances, they do not have access to negative evidence, and hence inevitably are unable to rule out overgeneral models of the language. "The child cannot use . . . absence as evidence, since so far as he or she is concerned the very next sentence could have [a positive example], and absence until then could have arisen from sampling error, or a paucity of opportunities for the adult to utter such sentences" (Pinker 1984: 48).

(b) *Efficiency*. This has two aspects. First, Pinker claims that there are too many possible distributional relations that are potentially relevant, and that exploring all these possibilities is combinatorially intractable. Second, he argues that distributional methods are liable to lead to inappropriate generalizations: "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: 49). More generally, Pinker argues that since pertinent linguistic generalizations are not couched in terms of simple distributional properties such as preceding word, first word in sentence, and so on, inappropriate generalizations are inevitable.

We shall argue that neither of these arguments applies to distributional methods to solve the bootstrapping problem for natural language. To show this we present a range of simulation results which show that considerable amounts of information about both syntactic categories, and the categories of *phrases*, can be derived using a distributional analysis of a large, noisy, unlabelled corpus of English.

The learnability argument

The learnability argument is that negative evidence is essential to rule out overgeneral models of the language if a purely distributional approach is taken. If valid, this argument would seem to have extremely disturbing consequences for the feasibility of induction in many domains, not just syntax. In particular, the whole of empirical science is built exclusively on "positive evidence". There is, after all, no oracle which tells the physicist, chemist or biologist what does information, rather than the empirical case for the importance of each in child language acquisition.

Possible sources of syntactic information

There are three 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, and on the analysis of prosody.

Distributional or correlational bootstrapping

Various authors (Kiss 1973, Maratsos 1979, Maratsos & Chalkley 1981, Maratsos 1988, Finch & Chater 1992) have suggested that words of the same category tend to have a large number of syntactic regularities in common. For example, Maratsos suggests that word roots which take the suffix "-ed" typically take the suffix "-s" and are verbs. Words which take the suffix "-s", but not the suffix "-ed" are typically count-nouns. Consequently, if we take a large number of predicates such as *takes the suffix "-s"*, *takes the suffix "-ed"*, *takes the suffix "-ing"*, *appears immediately after "the"*, and so on, there will be strong correlations evident. These correlations can be used, through some statistical analysis, to find proto-word classes which can later be refined to word classes more consonant with a mature language theory. Various other approaches, based on measuring local statistics of large corpora of language, have also been proposed (Brill et al. 1990, Marcus 1991, Finch & Chater 1992, Finch & Chater 1993, Schütze 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). 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 which the learner can apply, and it is possible, though not necessary, that these constraints might be specific to the task of acquiring language. So distributional methods may, in some sense, embody prior knowledge.

Semantic bootstrapping

Grimshaw (1981) and Pinker (1984, 1987, 1989) hold that the mechanism for the initial classification of words makes use of a correlation between syntactic information and prior semantic categories provided by evolution or learning. This account presupposes concepts such as *possession*, *action*, *objecthood* and so on, in explaining the early acquisition of syntactic categories. They can also assume that complex conceptual representations already exist of external events, and dependencies between these representations and the sound stream can be exploited to infer low-level syntactic structure. Thus, since there is a strong correlation between, for example, being an object and being referred to by a noun, semantic categories, which might be expected to be innately present in descriptions of the world, need only be correlated with the speech sound stream in order to infer rough approximations to a mature syntactic classification. Also, the concept of *noun phrase* might be semantically bootstrapped by defining it to be "that which refers to an object", together with some innate assumptions about the relationship between language and the extant mental representations. These rough approximations can then be further subjected to various forms of semantic and distributional analysis in order to refine them to be consonant with a more mature linguistic theory.

Another, somewhat different approach which also stresses the importance of extralinguistic context is what Curtiss (1987) terms the "social interaction" model (Snow 1972, Bruner 1975, Nelson 1977, Snow 1988). This approach stresses the child's communicative intent and the importance of the development of appropriate communicative relationships with care-givers. The pragmatic purpose to which language can be put by the learner, or by care-givers, is thought to crucially affect the course of acquisition. Thus, nouns can be thought of as words which can be used to denote agents and patients, and verbs as words which can be used to denote actions, etc. (e.g. Schlesinger 1971, Braine 1976, Schlesinger 1988).

Prosodic bootstrapping

Morgan & Newport (1981) propose that learners exploit the mutual predictability between the syntactic phrasing of a sentence, and its prosody. Consequently, if the child takes note of how something is said, he or she has information about the "hidden" syntactic phrasing of the sentence that the child needs to find for a mature theory of language. Thus the syntactic structure of language is not so well "hidden" after all, and an approximation to it may be found by listening to how a sentence is spoken.

Assessing the potential contributions of information sources

In order to quantitively investigate the amount of information that can 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 *not* happen; all that the scientist can do is observe what *does* happen (which is not the same every time a phenomenon is observed), and attempt to account for the data as well as possible. Thus, according to Pinker's account of distributional models, the language learner and the scientist are in just the same predicament. For both, it is never possible to definitively conclude that a phenomenon can be ruled out by distributional methods alone – the fact that it has not so far occurred may indeed have arisen from sampling error, or the like. The manifest possibility of scientific enquiry suggests that the learnability argument cannot be valid, either in general, or in the case of language learning.

Specifically, the problem with the learnability argument is that it does not take account of the fundamentally statistical character of inductive inference (whether these statistics are computed explicitly, or judged intuitively by the learner). Inductive inference involves choosing a model on the basis of a finite amount of data; it is not possible to find a model which is known to be correct, because there is always the possibility of later falsification, but it is possible to choose the model which is most probable, given the available data (using Bayesian statistical methods), to choose the model which makes the data most likely (using maximum likelihood methods), or to use some other criterion. Overgeneral models, which Pinker assumes cannot be ruled out without negative evidence, are rejected as highly improbable, since they predict the possibility of (classes of) data which are never observed (for a detailed discussion of inductive inference within a Bayesian model comparison framework see, for example, Earman (1992)). Pinker correctly describes methods which use the non-occurrence of tokens in a corpus as negative evidence as being dependent on the learning mechanism used, and therefore hard to evaluate, but does not go on to conclude that since the child certainly does have a learning mechanism, that it might well make use of non-occurrence as negative evidence.

For example, to return to Pinker's "slowly" example above, the use of distributional analysis might indeed derive the acceptability of "the slowly is good" from "John eats meat", "John eats slowly" and "the meat is good". However, empirically (in terms of the analysis that we shall describe below), the sequence "DET ADVERB-1 IS" is about 70 times less likely to appear than one would expect from chance if language was a random stream with lexical items appearing in proportion to how they actually appear. Here, "ADVERB-1" is the class of adverbs which includes "slowly", and "IS" is a class which includes "is, was, are, were, has, have". The non-appearance of this sequence is indicative of a syntactic constraint. Consequently, the non-occurrence of a sequence in a corpus can falsify (or make much less likely) a trivial hypothetical grammar.

Distributional methods can be shown to work

Although there may be no reason why distributional methods should not work in principle, Pinker's argument that they would be impracticable has yet to be addressed. The best way to answer this point is to provide a counter-example, where significant syntactic structure is demonstrably uncovered by linguistically naive distributional methods.

Recall that Pinker's main efficiency criticism is that relevant distributional statistics (e.g. subject/object relationships, head modifier relationships, etc.) are difficult to find, and that relationships which are easy to find (e.g. word adjacency) do not embody linguistic structure in a meaningful way, and consequently cannot be used to discover it. While we accept that highly linguistically relevant relationships are hard to find initially, we dispute the claim that simple relationships such as word adjacency cannot be exploited to find structure. Moreover, we shall show that the simple word-adjacency relationships can be used to infer much more linguistically perspicuous relationships encapsulating phrasal linguistic units of just the type which Pinker claims are most useful in discovering structure in natural language.

Finch & Chater (1992, 1993) proposed a tentative solution to the bootstrapping problem using distributional methods similar to that proposed by Kiss (1973). Kiss used a "most frequent first" approach, where the most frequent words appearing in a large corpus were clustered according to the similarity of statistical measurements of the lexical contexts in which they featured. This is in line with the view that it is not initially necessary to provide a theory which accounts for the acquisition of all of natural language in order to solve the bootstrapping problem, but rather just a significant part of it. The relations "last word", "next word", "last word but one" and "next word but one" were used as the basis of this classification. Although the methods used were not those proposed by Maratsos & Chalkley (1981), the spirit of the enterprise is similar – find some relationships which are highly correlated with syntactic structure, and use these to infer a syntactic classification for words. It was found that for the most frequent 2000 words, a highly linguistically perspicuous classification was uncovered, which featured all of the main word classes.

Pinker argues that one of the main problems with the efficiency of distributional bootstrapping is that there are potentially a very large number of distributional relationships which can be used to uncover linguistic structure.

> Perhaps, then, one can constrain the child to test for correlations only among linguistically relevant properties. There are two problems with this move. First of all, most linguistically relevant properties are abstract [e.g. syntactic categories, grammatical relations] ([this argument] owes its force to the fact that the contrapositive (roughly) is true: the properties that the child can detect in the input – such as serial positions and adjacency and co-occurrence relations among words – are in general linguistically irrelevant). (Pinker 1984: 49–50)

It may be true that the learner cannot formulate distributional generalizations in terms of linguistic abstractions, at least in the early stages of acquisition when presumably linguistic input cannot be parsed in appropriate linguistic terms.

Even if the relevant linguistic abstractions were available, then the results of distributional analysis would still only approximately specify the syntactic categories of individual lexical items (for instance, distributional tests in linguistics are useful heuristics for, rather than litmus tests of, category membership) (Radford 1988). For an information source to be useful, however, it does not have to be unequivocal. What is required is only that it is reliably statistically correlated with relevant linguistic regularities in real speech. Many perceptible relationships in the linguistic input, indeed the very examples that Pinker cites, have been shown to satisfy this requirement (regarding adjacency and co-occurrence relations, see Finch & Chater (1992), Finch (1993) and Schütze (1993); regarding serial position in sentences, see, for example, Hughes (1992)). Below we report our own recent work applying distributional methods to learning the syntactic categories of phrases, rather than just individual lexical items.

Finding phrasal categories

The rest of this chapter addresses the problem of uncovering syntactic structure at a higher level than just word classes. According to the standard view, the relevant level of linguistic analysis is a phrase-based one, where phrases are structured into trees, and are assigned labels, such as *noun phrase*, *prepositional phrase*, *sentence* and the like. We consider the degree to which it is possible to infer classes for *sequences* of words, as has previously been shown for word classes.

First, an initial classification of words is derived, and this classification is exploited to derive a classification of short (one-, two- and three-word) phrases. Then this classification is used to derive a syntactic classification of longer phrases.

Finch & Chater (1992) showed how a distributional analysis could roughly find syntactic categories.[2] They compiled a contingency table of 2000 common words against the contexts in which they appeared in a 40 million word corpus of USENET newsgroup articles. The context was simply defined to be the preceding two and following two words. To keep the computations tractable, attention was restricted to context words which were among the 150 most common words observed in the corpus. The context we used can therefore be thought of as four vectors of 150 dimensions, each dimension corresponding to one of the 150 most common words. The value of the vector is then given by the number of times the focal word appeared in the relevant relation (i.e. preceding, following, last but one, next but one). A definition of similarity between observed distributions of contexts was given (the Spearman rank correlation coefficient), and a cluster analysis performed to produce a hierarchical ontology of the words.

By stopping the hierarchical cluster analysis after only a certain number of links have been made, it is possible to find many classifications of words (i.e. partitions of the 2000-item word set). We stopped the classification when 500 categories remained (i.e. 75% of links were made), and chose the 100 most common of these as a classification of the "frequent part" of natural language. Our choice of these values is *post hoc* – not all values give such good results. In particular, if a very small number of categories is allowed, the distinctions between them have no obvious linguistic meaning. Nonetheless, linguistically meaningful results are obtained over a wide range of parameter values. Nearly all of these categories corresponded to linguistically coherent categories or subclasses of categories. For instance, of the 100 categories, the two most common were (see Finch & Chater (1992, 1993) for more detailed results):

(a) **C1** the my your their his our its a an any some several another every these those such each no many most certain

(b) **C2** of in on at for with from by into through against about between without under within during via upon towards toward across among beyond regarding

The corpus can now be mapped from a sequence of lexical items to a sequence of what we call *C-level* categories. For example, every occurrence of "the" would be replaced by "C1". Sequences of length 1, 2 and 3 of these *C-level* categories were searched for in a large corpus, and the 3000 most common such sequences were chosen for distributional analysis. This time the context was defined to be the four surrounding *categories* rather than the four surrounding words. Again a cluster analysis was performed, and again this was terminated when 75% of the links had been made, resulting in a classification of *short sequences* (X1, X2, . . . , X150). Several of these *X-level* short sequences had interesting linguistic interpretations. For instance, one, which contained about 80 short sequences, seemed to correspond to short noun phrases, in that in new text they corresponded to word sequences such as "it", "the" "apparent size", "each article", "the mother", "the real data", "a scientific theory". Note that all the examples given here were randomly sampled from a corpus of USENET articles which were not included in the corpus used for categorization. There is a preference towards longer examples, mainly to avoid repetition. Another category corresponded to parts of the verb "to be", including as exemplars "has been", "will have been", "is", "are", "might be" and so on. Other linguistically perspicuous classes include prepositional phrases, *n*-bar phrases and parts of the verb "to have". There are many linguistically imperspicuous categories, however, but many of these correspond to apparently coherent classes, even though most linguists would not use them. For instance, one class includes "the top of", "the name of", "the person with" and so on. Another one, which was picked at random, includes "use the", "use at the", "break into these", "add an". This is not a perspicuous category, but if a noun phrase lacking a determiner is added, it becomes a simple verb phrase. This observation suggests the utility of a further stage of analysis, in which sequences of *these* categories are clustered together to find still higher-level structure.

Sequences of these X-level *short sequences* of length 1 and 2 were searched for in a corpus of 40 million words taken from USENET newsgroups (stripped of headers, footers and repetition). The 3000 most common of these sequences were chosen for analysis, and this time the context was the set of the surrounding four X-level categories. Since there are many ways to parse a stream of words into X-level categories, each focal sequence can have many different contexts associated with it (as opposed to one for the procedure above). For example, the sequence "the big black dog" can be parsed in many ways. In particular, if each constituent of the parse is to be a short sequence (of length 1, 2 or 3), this phrase can be represented by labelled bracketings of short sequences in seven ways: (X81: The) (X32: big) (X32: black) (X36: dog); (X81: The big) (X32: black) (X36: dog); (X81: the) (X32: big black) (X36: dog); (X81: the) (X32: big) (X36: black dog); (X81: The big black) (X36: dog); (X81: the) (X36: big black dog); (X81: The big) (X36: black dog). Each labelled bracketing, or *parse*, corresponds to a sequence of X-level categories: in this case, the sequences are X81 X32 X32 X36; X81 X32 X36; X81 X32 X36; X81 X32 X36; X81 X36; X81 X36; X81 X36.

If "the big black dog" was the left context of an item of interest, then although the immediately preceding category is always X36, the last but one category is either X81 (noun premodifier with determiner) or X32 (noun premodifier without determiner).

The method was applied to a corpus of 10 million words, and again a cluster analysis was performed and terminated when 75% of the links had been made, leaving 135 classes. Five of the eight most frequent classes correspond to coherent linguistic entities. The others are coherent, but end in determiners, and so are not classical constituents (although they would be categories in a categorial grammar). Table 12.1 shows some examples of word sequences found to be in these five phrasal classes. We also give a small selection of random sequences from the corpus to show that the elicited categories really do uncover significant structure relative to random selection. We label with "*" those sequences which could not be analyzed as their description by a linguist, and by "?" those which are strange. In parentheses we give the percentage of un-starred examples.

As can be seen from Table 12.1, the classification is not entirely accurate, but remember that our goal is not to find a correct classification of language immediately, but rather to find significant amounts of structure which can later be refined by other methods which might make use of semantic and prosodic information. In many of the classes, over 80% of members could have the same syntactic category (recall that our aim is not to "parse" sentences, but rather to find what structure might be posited as a plausible arc by, for example, a chart-parser). Thus, it seems that distributional methods can provide significant information concerning phrasal level syntactic structure, even when used in isolation.

The important point is that some non-trivial structure of language has been learned, and that this has been done by applying non-language-specific distributional techniques to raw language data. For information, in a corpus of 500 000 valid words (i.e. words in the most frequent 2000 words in the corpus), 1 730 000 constituents were discovered. Of these, 1 400 000 were coherent

Table 12.1 This table shows some token sequences of four of the classes found by empirically clustering word sequences according to the similarity of their contexts of occurrence.

Simple sentences: what is a context, that's a different story, you will also receive a copy, we could hold some events, you must continue, we have the chance, some groups have no names, you start out, * you have any problems, the project should work, the old version is still available, ? I think it, I will have the car, you are standing, there's always the chance, I kept them, it would be appropriate, I think there's a piece, there is a french culture office, ? I would argue, * it is called, the bar could be seen, it's ok, the conference is over, I was talking to a friend. (92%)

Verb phrases: give away, pick them up, buy some audio tapes, suggest a company, have a new book and manual, get away from it, ask them to change the entry, change the entry, * think the world, can't remember what day, got nothing, disagree, do something, look around for people, go, even understand the questions, really want an argument, be appropriate, need to move out, get the information, try to send them, know of a place, live, read about them, don't have my copy, get to the question, get back on this, ? tell this, make them, go around, change the subject, know it, call the previous owner, give the name of their version, * believe that their version, can't see anything, have to have messages. (97%)

Noun phrases: ? the situation theory and its applications, the natural language group, some sort of code, * this since it, a new reference to the database, their hands, ? the bar with their parents, that day, the logical structure of natural languages, * me for a game, * it on line, a case, some of my stuff, a change of date, what parts, the rights to them, the number one, the end of the world, ? the money on a government, * the name of product, something similar, a fairly normal life, ? a dog to the club, * the point where it, what number, * some areas this, that way, the attention, * that names, * that names and references, many of the good responses, ? several friends in this, several friends in this area, any of the above equipment, ? another in your opinion. (77%)

Infinitival complements: to accept this attitude, * to allow laser printer, to be about her, to be at an end, to buy more, to call them, to change the name, to come up, to find the problem, to get a piece, to get me back, to get over it, to have a baby, to hear more, to keep it, to leave an engine, to mention me, to mention the groups in question, to pay the high prices, to play them, to read in the shell window, ? to replace the include, to run their own bbs, to start, to start a discussion, to take it, to take over the world, to take this out, to use the drive, to use the old mode, to wait for the music, to write the software, to have brought it. (97%)

Prepositional phrases: of this network, for the family, in this, ? between the state, in the story, in your question, of the story, back to the list, of this article, with a person, in our state, in an order, in our culture, with the image, of the country, on the net, on this, to the parents, of india, at the door, in certain parts, in the same area, * in article, for the us, ? in the general, with the local party, for the community, in the original post, on the individual, on engineering, of the free world, in those countries, to the other states, by the indian army, on men, * by one of her, by an individual, of the world, on a host, ? in the history, down in the country, of ancient times, of small discussion, from him, on his relationship, of news, via this, in india, * without star. (94%)

Random sequences: whether such political rubbish should temporarily as visitors, issue on, the involvement of subhash, this message, looks like because of, that is dharma from, to the, so i checked that too, my situation was, why those who keep their, information about i feel you, problem is that, if others, who promote and protect the, shouldn't you reserve judgement, it is a, islam was categories, where at least 80% of their member tokens were considered possible to have the same simple category in categorial grammar. Of the word tokens, 90% were in some category of length at least 2, and 80% of these were coherent categories. The categories listed above covered 35% of the corpus (i.e. 35% of all word tokens appeared in at least one of the above categories). This figure increases to 65% if another two highly coherent categories (noun groups and verb phrases) are included.

Conclusions

Distributional methods have been shown to be able to uncover significant linguistic structure at several levels in natural language. In particular, we have demonstrated the relative ease of distributionally bootstrapping abstract linguistic entities including approximations to all word classes, relatively simple noun phrases, verb phrases, prepositional phrases and sentences. Although much "fine-grain" structure in natural language, such as verb subcategorization frames, has not been demonstrated, it is plausible that more sophisticated distributional methods will be capable of finding more subtle regularities. Indeed, some verb subcategorization information has been acquired, since although "disagree" is classified as a verb phrase, other single verbs such as "do" or "buy" classified as simple verb phrases only if followed by a candidate object.

This work suggests a number of interesting avenues for future research. These methods can be applied to corpora which more accurately reflect the input received by the child. The CHILDES database (MacWhinney & Snow 1985) provides over 2 million words of transcribed care-giver speech, which, while large enough to find an initial classification of words (see Redington et al. 1993), is too small to apply the techniques described here in full.

Another interesting question concerns the applicability of these methods to languages other than English. Whilst many of the grammatical regularities indicated in English by word order are, in other languages, more reliably indicated by various morphological regularities (e.g. case marking and so on), there is no reason why the general method described here should be restricted to exploiting word sequence regularities. Other sources of distributional regularity, such as inflectional ending, morphological structure and so on, might be exploited to derive syntactic information. However, it should also be noted that even for languages which do not have mandatory word order constraints, word order is still probably highly informative of syntactic category, so the methods used here may work well even with these languages. Both of these research avenues are currently impeded by the paucity of very large machine readable corpora.

The success of distributional methods in discovering syntactic categories at the lexical and phrasal level raises the question of the scope of such methods in other areas of language acquisition, including the acquisition of grammatical rules. The general problem of language acquisition appears so difficult, and to be

solved so effortlessly by the child, that we suspect that many sources of information, possibly including an innate universal grammar (Chomsky 1980), may be involved.

Notes

1. Pinker allows that distributional analysis may have some role in language acquisition, when supplemented by other, more important sources of information – in particular, semantic information. While, as noted above, we suspect that it is highly plausible that information is integrated in this way in child language acquisition, we argue here that distributional information can be a surprisingly valuable source of information even when considered in isolation.
2. Note that this method assigns a single syntactic category to each lexical item. Since many lexical items are syntactically ambiguous, the challenge of capturing all possible readings remains. This is an important topic for further research, worked on, for example, by Kupiec (1993).

Acknowledgements

Portions of this chapter are based on Finch & Chater (1994).

References

Braine, M. D. S. 1976. Children's first word combinations. *Monographs of the Society for Research in Child Development* **41**, 25–67.

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

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

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

Chomsky, N. 1980. *Rules and representations*. Boston, Mass.: MIT press.

Earman, J. 1992. *Bayes or bust*. Cambridge, Mass.: Bradford Books/MIT Press.

Finch, S. 1993. *Finding structure in language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Finch, S. P. & N. Chater 1992. Bootstrapping syntactic categories. *14th Annual Conference of the Cognitive Science Society, Proceedings*, 820–25. Hillsdale, New Jersey: Lawrence Erlbaum.

Finch, S. P. & N. Chater 1993. Learning syntactic categories: a statistical approach. In *Neurodynamics and psychology*, M. Oaksford & G. D. A Brown (eds), 295–322. London: Academic Press.

Finch, S. P. & N. Chater 1994. Distributional bootstrapping: from word class to protosentence. *16th Annual Meeting of the Cognitive Science Society, Proceedings*, 301–6. Hillsdale, New Jersey: Lawrence Erlbaum.

Grimshaw, J. 1981. Form, function, and the language acquisition device. In *The logical* *problem of language acquisition*, C. L. Baker & J. McCarthy (eds). Cambridge, Mass.: MIT Press.

Hughes, J. 1992. The statistical inference of parts of speech. Unpublished paper, Department of Computer Science, University of Lancaster.

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

Kupiec, J. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. *31st Annual Meeting of the Association of Computational Linguists, Proceedings*, 17–22.

MacWhinney, B. & C. Snow 1985. The child language data exchange system. *Journal of Child Language* **12**, 271–95.

Maratsos, M. 1979. How to get from words to sentences. In *Perspectives in psycholinguistics*, D. Aaronson & R. Rieber (eds). Hillsdale, New Jersey: Lawrence Erlbaum.

Maratsos, M. 1988. The acquisition of formal word classes. In *Categories and processes in language acquisition*, Y. Levy, I. M. Schlesinger, M. D. S. Braine (eds), 31–44. Hillsdale, New Jersey: Lawrence Erlbaum.

Maratsos, M. & M. Chalkley 1981. The internal language of children's syntax. In *Children's language*, vol. 2, K. E. Nelson (ed.). New York: Gardner Press.

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

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

Nelson, K. 1977. Facilitating children's syntax acquisition. *Developmental Psychology* **13**, 101–7.

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

Pinker, S. 1987. The bootstrapping problem in language acquisition. In *Mechanisms of language acquisition*, B. MacWhinney (ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Pinker, S. 1989. *Learnability and cognition*. Cambridge, Mass.: MIT Press.

Radford, A. 1988. *Transformational grammar*, 2nd edn. Cambridge: Cambridge University Press.

Redington, F. M., N. Chater, S. Finch 1993. Distributional information and the acquisition of linguistic categories: a statistical approach. *15th Meeting of the Cognitive Science Society, Proceedings*, 48–53. Hillsdale, New Jersey: Lawrence Erlbaum

Schlesinger, I. M. 1971. Production of utterances and language acquisition. In *The ontogenesis of grammar*, D. I. Slobin (ed.). New York: Academic Press.

Schlesinger, I. M. 1988. The origin of relational categories. In *Categories and processes in language acquisition*, Y. Levy, I. M. Schlesinger, M. D. S. Braine (eds). Hillsdale, New Jersey: Lawrence Erlbaum.

Schütze, H. 1993. Word space. In *Advances in neural information processing systems 5*, S. J. Hanson, J. D. Cowan, C. L. Giles (eds). San Mateo, Calif.: Morgan Kaufmann.

Snow, C. 1972. Mother's speech to children learning language. *Child Development* **43**, 549–65.

Snow, C. E. 1988. The last word: questions about the emergence of words. In *The emergent lexicon,* M. Smith and J. Locke (eds). New York: Academic Press.

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