Distributional Bootstrapping: From Word Class to Proto-Sentence

Steven Finch

Human Communications Research Centre
University of Edinburgh
2, Buccleuch Place
Edinburgh, U.K.
EH8 9LW

+44 31–650–4656

steve@cogsci.ed.ac.uk

Nick Chater

Department of Psychology
University of Edinburgh
7, George Square
Edinburgh, U.K.
EH8 9JZ

+44 31–650–3432

nicholas@cogsci.ed.ac.uk

Abstract

There have been various suggestions about how children might acquire a proto-classification of elements of natural language, such as is conjectured to be necessary to allow the child to "bootstrap" language acquisition (Maratsos 1979; Pinker 1984). One, proposed by Kiss (1972) and Maratsos (1979), but criticised by Pinker (1984), is that children look for distributional correlations between simple linguistic phenomena in the language they hear in order to derive more sophisticated abstract linguistic classifications. Finch & Chater (1992) showed that a relatively complete syntactic classification of the lexicon could be found for common words in natural language using distributional bootstrapping.

This paper reviews some of the arguments Pinker raises against distributional methods, and then describes a system which overcomes his objections, where sequences of words are classified into phrasal classes by a linguistically naive statistical analysis of distributional regularities from a large, noisy, untagged corpus. For many classes, such as *sentence* and *verb phrase*, the accuracy of the classification (ie. the proportion of putative sentences which can in fact be linguistically interpreted as sentences) is in the region of 90%, thus enabling the child to break the "bootstrapping problem".

Introduction

Acquiring syntax appears to face the child 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 child appears to be faced with what seems an impossible task: searching the entire space of category/rule combinations simultaneously. Three broad approaches to the bootstrapping problem can be distinguished.

Distributional or *correlational bootstrapping* (Maratsos 1979; Maratsos & Chalkley 1981; Finch & Chater 1992) makes use of the fact that words of the same category tend to have a large number of syntactic regularities in common. For example, 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 (in the light of interpretation within an increasingly complex and accurate, adult-like theory) to word classes more consonant with a mature language theory.

Semantic bootstrapping (Pinker 1984, 1988) holds that the mechanism for the initial formal classification of words makes use mainly of the structures that evolution and learning have given the child in interpreting the external world. Consequently, such theories can refer to 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 semantically 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, these 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 linguistic 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.

Prosodic bootstrapping (Morgan & Newport 1981) exploits the mutual predictability between the syntactic phrasing of a sentence, and the way it is said (ie. 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 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 may be easily found by listening to how a sentence is spoken.

One might expect a child to make use of as much information as possible in acquiring a theory of language, and as such one might expect the child to use all these

sources of information. Nonetheless, some of the sources may easier to exploit, at least initially, than others. For example, in order to exploit semantic regularities, one needs to be sure that the words which are spoken are connected with mental representations of the situation the child finds themself located in. Since adult speech includes a large number of references to abstract entities, and events which occur outside the immediate experience of the child seeking to acquire language, such regularities which do exist between language and mental representations will be clouded by the "noise" of language not being used to refer to mentally represented events. On the other hand, both prosodic and distributional information is (relatively) explicit in the sound stream itself, so this information may be used to infer linguistic structure from all adult speech, whatever its subject matter.

Pinker's critique of distributional methods

Pinker (1984, 1987) argues that distributional methods, and in particular the approach of Maratsos & Chalkley (1981), suffer two fundamental problems when used in isolation (he suggests that they must be augmented with additional, semantic information). 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.

Learnability. Since distribution 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)

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 apply to distributional methods to solve the bootstrapping problem for natural language, and present a range of simulation results which show that considerable amounts of information about both syntactic categories, and the the categories of *phrases*, can be derived using a distributional analysis of a large, noisy, unlabeled corpus of English.

The *learnability* argument

The learnability argument is that negative evidence is essential to rule out overgeneral models of the language. 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 *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 that data as well as possible. Thus, according to Pinker's account, 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 — 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. Over-general 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.[1] 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 naive methods in 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, 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

[1] for a detailed discussion of inductive inference within a Bayesian model comparison framework, see Earman 1992.

of a sequence in a corpus can falsify (or makes much less likely) a trivial hypothetical grammar.

Distributional methods can be efficient

Although there may be no reason that 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 counterexample, where significant syntactic structure is demonstrably uncovered by linguistically naive distributional methods.

In Finch & Chater (1992), we proposed a tentative solution to the bootstrapping problem using distributional methods similar to that proposed by Kiss (1972). It was 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. It remained to extend this scheme so that higher-level syntactic structure could be uncovered.

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. This may be true, but it would seem that very simple ones, only involving frequent adjacent words, suffice to uncover a good approximation to word classes. Since it is entirely possible that a child could have innate knowledge about where to look for linguistic regularities, rather than about the precise nature of these regularities as is the case with a "universal grammar", demonstrating the utility of such simple but informative relationships for distributional bootstrapping suggests that although it may be true in general that distributional bootstrapping is hard for the reasons which Pinker argues, the abstract classes in natural language lend themselves to discovery by this kind of analysis.

> 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 [eg. syntactic categories, grammatical relations etc.] ([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 p49–50**

While true in general, for distributional techniques to work the relationship need only be statistically relevant, in that the relation is reliably statistically correlated with relevant linguistic regularities in real speech. This is true for many perceptible relationships (eg. adjacency and co-occurrence relations (Finch & Chater 1992; Finch 1993; Schultze 1993), serial position in sentences (eg. Hughes 1992), and probably many more).

The rest of this paper 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. Is it possible to infer classes for *sequences* of words in much the same way as we did for word classes? If this is the case, this opens the possibility that a larger part of natural language can be bootstrapped by an unsupervised, non language-specific learning mechanism than has previously been demonstrated, and possibly reduces the amount of innate knowledge which is needed to acquire language. In order to show the feasibility of using unsupervised methods to acquire natural language, we present a three stage hierarchical analysis of language. First, an initial classification of words is derived, and this classification is exploited to derive a classification of short (1, 2, and 3 word) phrases. Then this classification is used to derive a syntactic classification of longer phrases. The next section describes this process.

Finding phrasal categories

In Finch & Chater (1992), we showed how a distributional analysis could roughly find syntactic categories. We collated a contingency table of 2000 common words against the contexts in which they appeared in a very large 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 (ie. partitions of the 2000 item word set). We stopped the classification when 500 categories remained, and chose the 100 most common of these as a classification of the "frequent part" of natural language. Nearly all of these categories corre-

sponded to linguistically coherent categories or subclasses of categories. Thus we ended up with 100 categories, the two most common of which were:

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

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 *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[2] *it, the apparent size, each article, the mother, the real data, a scientific theory.* 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 isn't a perspicuous category, but if a noun phrase lacking a determiner is added, it becomes a simple verb phrase. This observation clearly suggests the next stage in the analysis, which is to cluster together sequences of *these* categories to find still higher level structure.

Sequences of these X-level *short sequences* of length 1 and 2 were searched for in a large corpus. 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 1 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 labeled bracketings of short sequences in 7 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 the 7 labeled 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 category of any parse of "the big black dog", the last but one category is either X81 (noun pre-modifier with determiner) or X32 (noun pre-modifier without determiner). Thus we can construct the contingency table of short phrases against their contexts, and define the similarity of two phrases to be some statistical measure of similarity between their contexts. For this, we used the Spearman rank correlation Coefficient, since it is known to be powerful in circumstances where linear correlation is not.

A corpus of about 10 million words was used to construct the contingency table, and again a single link cluster analysis was performed and terminated when 75% of the links had been made. The examples given are randomly sampled from a parse of some USENET newsgroup articles which were not part of the corpus used to infer the classification. There is a bias towards longer exemplars, mainly to avoid repetition of short exemplars. Figure 1 shows some examples of word sequences found to be in four of the phrasal classes which were the result of this operation.[3]

As can be seen from this figure, the classification os 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. Many of the classes are over 90% accurate. Consequently, this constitutes significant evidence against the assertion that it is not possible to efficiently infer significant linguistic structure from distributional techniques alone.

Conclusions

Distributional methods have been demonstrated to be quite powerful at uncovering significant linguistic structure at many levels in natural language, and, moreover, have been able to do this without exploiting sophisticated knowledge about the nature of the regularities present in natural language. In particular, we have demonstrated the relative ease of distributionally bootstrap-

[2] All the examples given are randomly taken from a text not included in the corpus used for categorisation. There is a preference towards longer examples (mainly to avoid repetition).

[3] This is a selection of four classes from 100 or so classes we could have chosen. Some of the classes were not so linguistically coherent. For example, one of the classes corresponded to subject+verb, which is not usually considered a constituent. In general, most of the categories would be considered largely coherent in a categorial or dependency grammar (Barry & Pickering 1992).

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.*

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.*

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 faily 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.*

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.*

Figure 1: This figure shows some token sequences of four of the phrasal classes found by empirically clustering word sequences according to the similarity of their contexts of occurrence.

ping 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 subcategorisation frames, has not been demonstrated, it is entirely plausible that more sophisticated variants of the methods described here will be capable of finding such regularities[4].

Of course, this work is very far from demonstrating the feasibility of the entirely unsupervised acquisition of natural language. In particular, it does not even address issues concerned with *parsing*, such as determining the phrase structure of a sentence, nor does it directly address the fundamental question of language learning: How can an agent learn a *generative grammar* from a corpus, rather than just a classification of some rather short phrases. These questions are the subject of ongoing research, as is the relationship between distributional and semantic bootstrapping. It would be surprising if both distributional and semantic criteria were not used by a child learning language, since both distributional and semantic regularities are potentially readily available to children, and both are informative of syntactic structure and hence can be expected to aid the child in their aquisition of syntax. The same is true for prosodic information.

Other questions for further research include demonstrating the efficacy of this approach using corpora of 'caregiver data' (see Redington, Chater & Finch, 1993), and demonstrating the portability of this method for uncovering structure to other languages. This research can continue when very large corpora from these domains become available. For example, the CHILDES database, at just over 2 million words, while large enough to find a rough initial classification of words, is too small to apply the techniques described here in full.

In other languages, many of the grammatical regularities indicated in English by word order are more reliably indicated by various morphological regularities (eg. case marking and so on). There is no reason why the general method described here should be restricted to exploiting word-sequence regularities, and other sources of regularity, such as inflectional ending, and so on, might be exploited to derive hierarchical structure. However, it should also be noted that even for these languages, word order is still probably highly informative of syntactic category, so the methods used here may work well even with these languages.

References

Baker, J. (1982) Trainable Grammars for Speech recognition. in D. Klatt & J. Wolf (eds.), *Speech Communication Papers for the 97th meeting of the Acoustical Society of America*, ASA, pp. 547 – 550.

Barry, G. & M. Pickering (1992) Dependency & Constituency

[4]Some verb subcategorisation information has indeed 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.

in Categorial Grammar. in *L'ordre des Mots dans les Grammaires Catégorielles*. A. Lecompte (Ed), Adosa, Clermont-Ferrand.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT press, Boston, Mass.

Earman, J. (1992) Bayes or Bust. Cambridge, MA: Bradford Books, MIT Press.

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

Finch, S. P. & N. Chater (1992) Bootstrapping Syntactic Categories. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America*. Bloomington, Indiana. 820–825.

Fong, S. & R. Berwick (1992) *Parsing English and Japanese with Principles and Parameters Theory*. Paper presented at the Fifth Annual CUNY Conference on Human Sentence Processing.

Fujisaki, T., F. Jelinek, J. Cocke, E. Black, T. Nishino (1988) Probabilistic Parsing Methods for Sentence Disambiguation. In *Proceedings of the International Parsing Technologies Workshop*. Carnegie-Mellon University, Pittsburgh.

Hughes, J. (1992) The Statistical Inference of Parts of Speech. *Manuscript, University of Lancaster, dept. of Computer Science, UK*.

Kiss, G. R. (1972) Grammatical Word Classes: A Learning process and its Simulation. *Psychology of Learning and Motivation* **7** 1–41.

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

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

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.

Pereira & Schabes (1992) Inside-Outside Reestimation from Partially Bracketed Corpora *Fifth DARPA Speech and Natural Language Workshop, February, 1992*

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

Pinker, S. (1987) The Bootstrapping Problem in Language Acquisition. In B. MacWhinney (ed.) *Mechanisms of language Acquisition*. Hillsdale: Erlbaum.

Redington, F. M., Chater, N. & Finch, S. (1993) Distributional Information and the Acquisition of Linguistic Categories: A Statistical Approach. Proceedings of the 15th Meeting of the Cognitive Science Society, Hillsdale, NJ: LEA.