Bootstrapping Syntactic Categories Using Statistical Methods

Steven Finch
Centre for Cognitive Science
University of Edinburgh

2, Buccleuch Place

Edinburgh, U.K.

steve@cogsci.ed.ac.uk

Nick Chater
Department of Psychology
University of Edinburgh

7 George Square

Edinburgh, U. K.

nicholas@cogsci.ed.ac.uk

Abstract

In learning the structure of a new domains, it appears necesssary to simultaneously discover an appropriate set of categories and a set of rules defined over them. We show how this *bootstrapping* problem may be solved in the case of learning syntactic categories, without making assumptions about the nature of linguistic rules. Each word is described by a vector of bigram statistics, which describe the distribution of local contexts in which it occurs; cluster analysis with respect to an appropriate similarity metric groups together words with similar distributions of contexts. Using large noisy untagged corpora of English, the resulting clusters are in good agreement with a standard linguistic analysis. A similar method is also applied to classify short sequences of words into phrasal syntactic categories. This statistical appraoch can be straightforwardly realised in a neural network, which finds syntactically interesting categories from real text, whereas the principal alternative network approach is limited to finding the categories in small artificial grammars. The general strategy, using simple statistics to find interesting categories without assumptions about the nature of the irrelevant rules defined over those categories, may be applicable to other domains.

Steven Finch is a PhD student at the Centre for Cognitive Science, Edinburgh, supervised by Nick Chater.

Bootstrapping Syntactic Categories Using Statistical Methods

Steven Finch & Nick Chater

Abstract

In learning the structure of a new domain, it appears necessary to simultaneously discover an appropriate set of categories and a set of rules defined over them. We show how this *bootstrapping* problem may be solved in the case of learning syntactic categories, without making assumptions about the nature of linguistic rules. Each word is described by a vector of bigram statistics, which describe the distribution of local contexts in which it occurs; cluster analysis with respect to an appropriate similarity metric groups together words with similar distributions of contexts. Using large noisy untagged corpora of English, the resulting clusters are in good agreement with a standard linguistic analysis. A similar method is also applied to classify short sequences of words into phrasal syntactic categories. This statistical approach can be straightforwardly realised in a neural network, which finds syntactically interesting categories from real text, whereas the principal alternative network approach is limited to finding the categories in small artificial grammars. The general strategy, using simple statistics to find interesting categories without assumptions about the nature of the irrelevant rules defined over those categories, may be applicable to other domains.

The Bootstrapping Problem

One reason why learning the structure of a domain without any prior knowledge is so difficult is that both an appropriate set of categories to describe the phenomena and the rules defined in terms of those categories must be learned from scratch. Thus the learner must solve a "bootstrapping" problem: the specification of a set of rules presupposes a set of categories, but the validity of a set of categories can only be assessed in the light of the utility of the set of rules that they support. *Prima facie*, at least, this implies that both rules and categories must somehow be derived together. However, the space of possible of rule/category combinations is so large that it seems unlikely that such an approach will be feasible for learning the structure of any but the simplest domains.

Although the focus here will be natural language, the bootstrapping problem arises in the context of learning about any new domain. For example, in learning some new subject, say elementary physics, learners must somehow acquire both the relevant concepts and the correct rules of inference defined over those. For example, learners must grasp the concepts of momentum, force and so on, as well as how these concepts may be manipulated and interrelated using the formal rules. The bootstrapping problem is acute since these two projects are thoroughly interdependent - understanding the concepts presupposes some understanding of the rules in which they figure, and the statement of the rules presupposes the concepts that they interrelate. In the terminology of the philosophy of science, the development of science requires both new *natural kinds* and new *scientific laws* relating those kinds together. Thus the bootstrapping problem is at the heart of the problem of theory change, both in scientific enquiry and in individual cognitive development.

Rather than attempt to tackle the bootstrapping problem in its full generality, we shall focus on the test case of learning syntax as an illustration of a particular way in which the bootstrapping problem may be overcome. In syntax learning the bootstrapping problem is to learn the set of syntactic categories and the syntactic rules defined over them. Most work on formal models of syntax acquisition does not encounter the bootstrapping problem, since the syntactic category of individual lexical items are taken as given, and the focus is on deriving the set of rules defined over these items (that is, the corpus used in learning is *tagged*). Even given this restriction, of course, the problem of rule induction is very difficult, and there are a number of formal results (Gold 1967; Pinker 1984; Osherson, Stob & Weinstein 1986) which suggest that constraints on possible linguistic rules must be innately specified. We pursue a parallel approach, using an untagged corpus, and tackling the bootstrapping problem directly. We give no prior information to the learner, and attempt to derive both the stock of syntactic categories

and the syntactic category of individual words from scratch.

The general strategy that we use is straightforward: we collect very simple statistics from of the data set, in the hope that a similarity measure defined in terms of these statistics will reflect useful underlying categories. We then derive a set of categories on the basis of their similarity with respect to these simple statistics. Despite the simplicity of these statistics in relation to the complexity of the rules of syntax of natural language, redundancy in the data means that the categories generated are close to the categories given by standard linguistic theory. Thus, the bootstrapping problem can be solved by inferring categories directly from simply, readily available statistics, without needing to make assumptions about the nature of the relevant rules.

Once these categories have been found, we can tag the previously untagged corpus, marking each word with its syntactic category, and to attempt to find rules defined over these categories. This can allow us to find a set of higher level phrasal categories defined over categories for words already derived. Thus a hierarchy of categories and rules can be derived by iterating this process. This method also promises to allow the revision of initial categorisation decisions, based on impoverished assumptions concerning the set of rules, in the light of the rules derived (we shall discuss this below). Below, we outline how this approach has been applied to learning aspects of the structure of natural language.

An Algorithm for Bootstrapping Syntactic Categories

In order to illustrate the above suggestions concerning how empirical measures of similarity can be exploited to solve the bootstrapping problem, we now derive a linguistic taxonomy which is remarkably close to the orthodox view of the various species of syntactic category. In order to achieve this, a measure of similarity between words and phrases inspired by the "replacement test" of theoretical linguistics was used.

Empirical Similarity and Numerical Taxonomy

In traditional linguistics, words and phrases are categorised into several standard linguistic categories: nouns, verbs, noun phrases, and so on. One justification for this taxonomy is afforded by a number of "distributional tests", which assume that words and phrases which are distributed similarly should receive similar linguistic categories. Probably the best known test is the "replacement test"(e.g. Radford 1988):

> 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.

In traditional linguistics, "distribution" is grounded in linguistic intuitions as to whether a purported sentence is syntactically 'well-formed'. In the present context such intuitions cannot, of course, be presupposed, but the replacement test can be made empirically relevant by operationalising it as follows:

> Statistical Replacement Test
>
> Has the word or phrase been observed to occur in a corpus in similar contexts to another word or phrase? If so, then these should be given similar linguistic categories.

It remains to give formal accounts of what constitutes the "context" in which a word or phrase appears, and to define some measure of "similarity" between two such contexts.

To avoid unnecessary presuppositions about the structure of language, we assume an extremely simple definition of the context of a word - the context is simply 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).

There were several candidates for this which were quite good at uncovering structure automatically. In the spirit of the statistical replacement test described above, we propose that any reasonable measure of similarity defined to elucidate linguistic distributional similarity should be insensitive to the absolute frequency of ocurrance of any particular word, but should be dependent on the position it is observed to occur at relative to other words. That is, it should satisfy the following criterion:

> **Replacement Criterion** If every occurrence of a word, w, is replaced throughout the whole corpus independently and at random by w' with probability p, and w'' with probability $1-p$, and neither w' nor w'' previously occurred in the corpus, then w' and w'' should have similar contextual distributions according to the chosen similarity metric.

A metric which gives hierarchical structure in accord with linguistic orthodoxy was found to be the Spearman Rank Correlation Coefficient between the vectors of frequencies of context words. Since Rank Correlation between two vectors of ranks is in the range $[-1, 1]$, we used an appropriate rescaling of values into the range $[0, 1]$.

Since Sokal & Sneath (1963) first introduced techniques of numerical taxonomy to the biological community, hierarchical cluster analysis has found a wide range of applications, especially in the biological and social sciences. We use our distributional similarity

metric as the basis for a hierarchical cluster analysis of words, which places words with similar distributions nearby in the hierarchy. Nodes in the resulting taxonomy correspond closely to traditional syntactic categories.

The goal, in the first instance, is to induce a standard syntactic categorisation. Then we analyse short phrases in a similar way to deduce similarities between phrases of various length, and thereby induce facts about the grammar describing them.

Computational Experiments

We have conducted a number of studies deriving syntactic categories from artificial data generated by a phrase structure grammar, and classifying letters and phonemes into linguistically interesting classes using corpora of real text (Finch & Chater 1991). Here we concentrate on the problem of finding syntactic categories in real corpora.

Syntactic categories in natural language

A 40,000,000 word corpus of USENET newsgroup data was stripped of headers, footers and the like. Even before cluster analysis, a list of the ten nearest neighbours of sample words shows that the Rank Correlation metric reveals at least some linguistic structure.

three: three, four, five, six, several, real, black, old, high, local, white.

I: I, we, they, he, she, you, I've, doesn't, don't, I'm, didn't.

south: south, east, west, north, war, public, government, tv, system, dead, school.

Clustering results

The tree structure for the entire set of words analysed, the 1000 most common words in the corpus, is much too large to display in a single diagram. Therefore, an overview of the structure of the tree is given, with labels a node corresponding to the predominant syntactic category of the items dominated by that node. A small number of items have no well defined syntactic category (for example, single letters of the alphabet and words connected with newsgroup administration such as "edu" and "com") and these were rejected from the analysis. Of the remainder, less than 5% are misclassified with respect to the label that we have given to their dominating node. Figure 1 therefore shows that the gross taxonomy of the lexical items is very close to a standard taxonomy of syntactic categories.

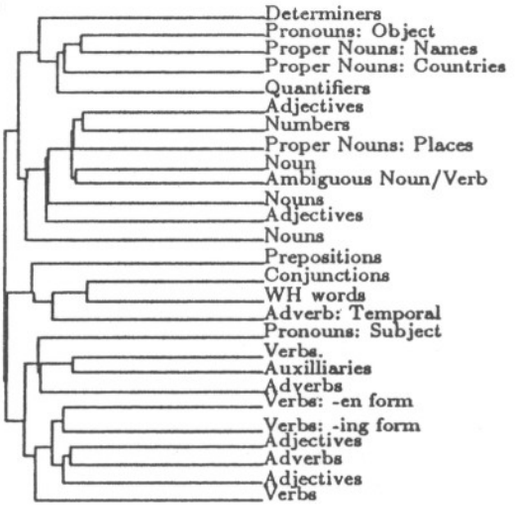

Figure 1

Figure 2(a) shows some of the low-level structure apparent within the whole dendrogram. The left hand dendrogram corresponds to part of the "adverbs" category of Figure 1. Note that some semantic regularities are apparent (really/actually, finally/eventually, thus/therefore, and so on). The other two dendrograms show respectively that low-level semantic features are revealed (being a computer term) and the dendrogram of subject-position pronouns shows a (relatively) orthodox syntactic analysis of pronoun/auxiliary contractions.

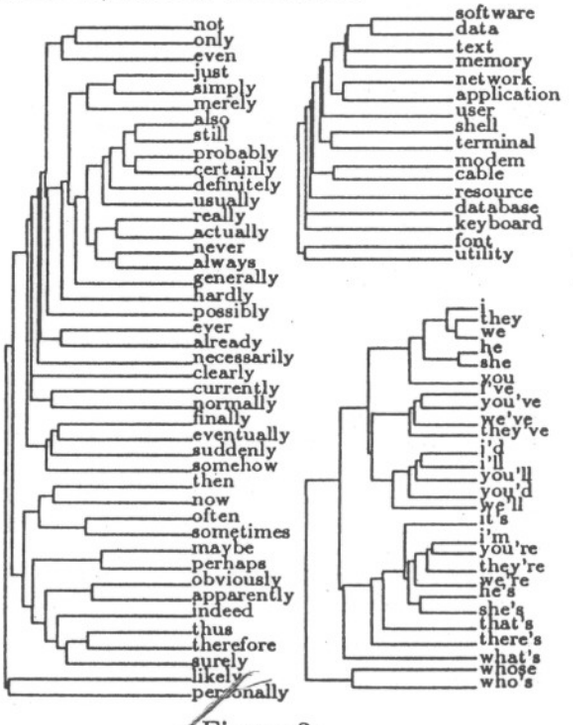

Figure 2a

Figure 2(b) shows low level structure for some adjectives, object position pronouns, countries, and numbers. Again it is clear that there is considerable accord between empirical and syntactic/semantic similarity.

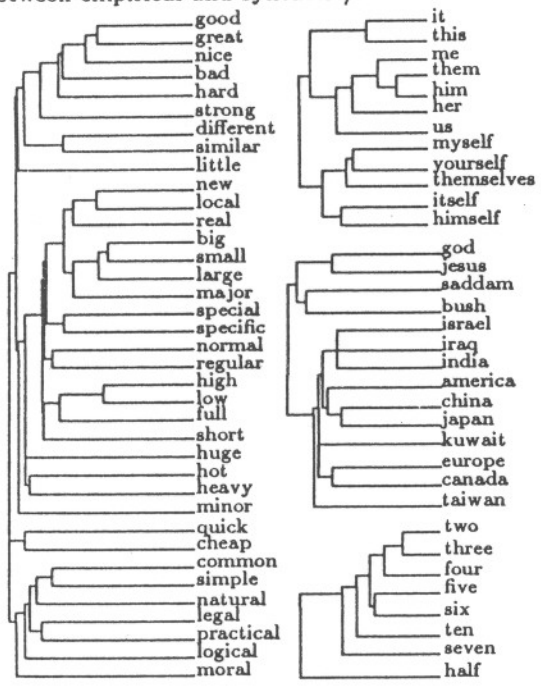

Figure 2b

Sequences and Similarity

After a hierarchical classification of lexical items has been derived, we can use this to classify sequences of categories hierarchically, and the derived similarity metric will again turn out to reveal interesting linguistic structure. This section details the experimental techniques and results of this analysis.

Classification The lexical hierarchy derived above was used to classify each lexical item by cutting the dendrogram at a particular level of dissimilarity, and thereby obtain several disjoint classes of words. Individual words were replaced with a code which corresponded to the class to which they belong, and the corpus was "parsed" accordingly. For instance, the two word sequences "the women", "the file" and "most data" were replaced by the sequence of labels "C30 C16". The principle advantage of this is one of sample size. If 600 words were in C16, and 20 in C30, for example, then the bigram "C30 C16" comprises, in principle, 12,000 word-level bigrams. This means that reliable statistics can be gathered on the "C30 C16" bigram with a much smaller corpus than needed for word-level bigrams. The situation is clearly exponentially worse for trigrams. For instance, the "C30 C16 C16" (Determiner Noun Noun) trigram corresponds to a possible 7,200,000 word-level trigrams. The size of the corpus is a major limitation to how far this unsupervised statistical approach can uncover the structure of language, and classification can be seen as a means of elucidating generalisations from (relatively) small corpora.

Results Rather than present dendrograms as we did for individual words, in order to show that interesting linguistic structure has been captured we instead show some of the "tightest" clusters. That is, the dendrogram is "cut" at a particular level of dissimilarity, and some of the resulting clusters are given as an illustration.

Noun Phrase Det Noun, Det Adjective Noun, Det Noun Noun, Det Verb/Noun, Det Adjective Verb/Noun, Det Inf, Det Verb/Noun Noun, Det Noun Verb/Noun, Det Inf Noun, Det ing Noun, Det PastPpl Noun, Det Det Noun, Det Adjective Noun, Det Adjective Inf, Det Adjective Verb/Noun, Det ing, Det Noun Adjective, Det Place Noun, Det Adjective QuantProNP

Note that the ambiguous category "Verb/Noun", which contains words judged to occur roughly equally frequently as non-finite verbs and nouns, behaves very much like "Noun" when preceded by a determiner. Even words which are typically non-finite verbs are judged similar to nouns when preceded by a determiner.

Verb Phrase Inf ProObj, Inf ProObj Noun, Inf Det Noun, Inf Det Verb/Noun, Inf Det Inf, Verb/Noun Det Noun, Verb/Noun ProObj, Inf ProObj Prep/Adv, Inf QuantNP, Inf QuantProNP, Inf ProObj Adjective, Inf Countries, Inf Noun, Inf Adjective Noun, Inf Noun Noun, Inf PastPpl, PastPpl PastPpl, PastPpl Adjective

Note that when followed by an object position pronoun, or a noun phrase, the ambiguous category "Verb/Noun" now behaves as (appears in the same contexts as) non-finite verbs.

Prepositional Phrase Prep Noun, Prep Det Noun, Prep Adjective Noun, Prep Det Verb/Noun, Prep Inf, Prep Det Inf, Prep Adjective Noun, Prep Verb/Noun, Prep Adjective, Prep QuantProNP, Prep ProObj Noun, Prep Conj&WH Noun, Prep Noun Noun, Prep QuantProNP Noun

Complex Nouns Noun
Noun, Noun, Noun Verb/Noun, Noun Preposition Noun, Noun Conj&WH Noun

Nouns are similarly distributed to compound nouns.

Auxiliaries Auxiliary Adverb, Auxiliary Adverb Adverb, Adverb Auxiliary, Auxiliary, Auxiliary TempAdvb, Auxiliary AdjMod, Auxiliary Adjective

As can be seen, auxiliaries can appear close to adverbs of various sorts, and the resulting phrase is similarly distributed to auxiliaries alone.

Relation to Neural Network Approaches

Outside the statistical tradition, there has been much interest in using neural networks to extract linguistic categories from raw data. In particular, Elman (1990, 1991; see also Chater 1989; Cleeremans, Servans-Schrieber & McClelland 1989) has shown how a recurrent neural network, trained to predict the next element in a sequence of inputs generated by a simple grammar, can develop patterns of hidden units which, when appropriately averaged and cluster analysed reveal underlying syntactic categories.

Elman's approach has a number of limitations. Firstly, it does not readily generalise to handle more realistic grammars, with many grammatical rules and a large lexicon. This is because the prediction tasks rapidly becomes extremely difficult, and learning is extremely inefficient and slow, if it occurs at all. Secondly, the linguistic categories are only implicit within the network, and can only be revealed using cluster analysis. However, cluster analysis on simple bigram statistics of the training corpus provide equally good clusters (Chater & Conkey in submission), so it is not clear how much statistical work the network is doing in uncovering the underlying linguistic categories.

The statistical analysis presented above suggests an alternative neural network approach, in which a network learns through simple Hebbian learning to represent words by their distributional context. Since similar words are assigned similar patterns, the network can find the relevant syntactic categories by performing a cluster analysis of the patterns. An attractive paradigm for unsupervised clustering is due to Kohonen (1982). This implements a variant of k-means clustering, where the k output units (or more exactly their weight vectors) correspond to the k-means which compete to account for portions of the data, to which they are most similar.

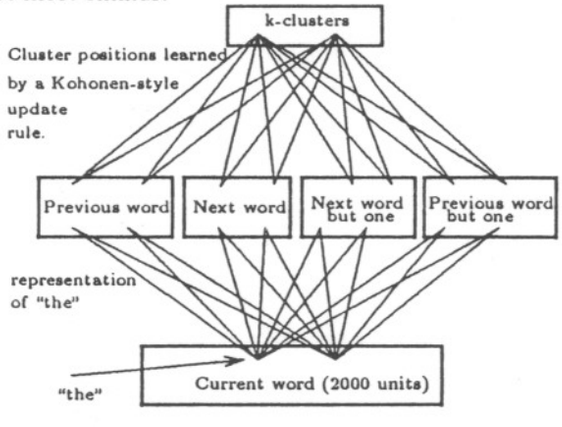

Figure 3

The network shown in Figure 3 corresponds to that used in our simulations with large corpora of real text. We used a similar, scaled down, version in the letter and phoneme level simulations reported below. The lower set of units use a localist representation of the current word (there are 2000 units, each corresponding to a different word under study). The middle set of units are divided into 4 banks, one bank corresponding to each of the four contextual bigram relations considered: last word but one, previous word, next word, next word but one. Only the most common 150 words were considered, and appearances of all other words in these contextual relations were ignored. The first layer of the net was trained with the 40,000,000 word newsgroup corpus simple Hebbian learning, with normalisation. After training, when a "current word" is presented, the middle layer represents the distribution of contexts in which that word occurs. The pattern representing each of the 2000 words are then clustered into 100 groups using a Kohonen network.

Network Simulations

First, a small network was given the task of clustering together letters, which were represented by the distribution of their surrounding context as described above for words. When the network consisted of two cluster nodes, it precisely divided vowels from consonants. The clusters resulting from a small (12,000 phoneme) corpus of phonemically transcribed speech (Svartvik & Quirk 1980), also approximately divided vowels from consonants as shown below.[1]

Vowels: @@ @ uu uh u oo o ng nd ii i e aa a

Consonants: zh z y @r w v th t sh s r p n m l k jh hi h g dh d ch b

In the word-level experiments, Some of the clusters obtained are shown below. In general words in the same cluster tend to have the same syntactic category, although there is sometimes more than one cluster which corresponds to the same syntactic category. Also some clusters appear to correspond to no linguistic category. Some of the clusters are shown below. Notice that one of clusters corresponds not to a single linguistic category, but consists of words which are ambiguous between two linguistic categories, nouns and verbs. In many of the categories there are one or two apparently spurious items, and some of the smaller categories, not shown, do not appear have any coherent linguistic basis. Although the categories are generally in accord with an orthodox syntactic classification, more linguistically perspicuous categories can be found by cutting the dendrogram produced in a full hierarchical cluster analysis at a particular dissimilarity level, to give disjoint clusters (as shown in Figure 1). Hence it may be possible to improve network performance further.

your those this these their the our one's my its his every each another an a

[1] We use the Machine-Readable Phonetic Alphabet.

why whom whether where what though that how because

two three ten six several half four five few fairly very

you've you're who's what's we're wasn't they've they're there's that's suddenly she's knowing it's i'm he's haven't comes being

washington v steve robert president peter mike michael math m john jesus japan iraq india george engineering david dave bell

yourself whatever us themselves them something someone somebody saddam myself me kuwait himself him her forth everyone anything

without within with when via unless under toward on near in if from for during by beyond between before at as among against across about

writing willing watching using turning trying thrown taking supporting showing sending selling seeing running putting printing playing paying passing making looking keeping giving getting flying finished finding doing considering coming changing calling buying behind acting

wanted used tried treated taught taken suggested stopped stated started sold shown seen saw saved responsible reported removed released received published provided produced presented posted played placed paid opposed noticed needed moved met looked led intended included heard found experienced done discussed died designed caught carried assumed associated asked applied allowed added accepted

window warning wall voice unit train track tape table stock statement stack signal screen sample role ring results ram purpose program process performance object months menu market map list link letter image ii frame format form foot flow filter film file faith entry effect dog distribution disks course contents chip box book article animal address addition account

walk wait use try stick sign share send save rid respond refer recognize reach protect pick pass offer occur miss keep judge include ignore hurt handle follow focus fix fill exist drop define count convert continue compile cause bring bother belong beat answer

words women views versions types tools tapes stories states sites responses questions programs products postings parents papers opinions numbers names movies laws ideas functions friends fonts fans experiences examples elements effects documentation discussions computers children cases canada applications advice

update transfer trade test split spell ride return report reply release register record present post plan move log lead force fly figure feed face escape end email die deal copy charge call break benefit attack

wonder wish win trust tell see say respect remember realize prove notice mention know imply imagine hope hear guess forget feel explain expect except doubt determine deny decide claim care blame believe assume ask argue agree

valid tough stupid somewhat slow simple silly separate related practical possible nice negative neat logical less intelligent important hot greater good faster expensive excellent easy correct closer blind better appropriate accurate

Discussion

We have shown how it is possible derive good approximations to the syntactic categories for English, without having a good account of the rules of syntax, by collecting statistics, deriving a similarity metric, and applying hierarchical cluster analysis. Further it was possible to use the lexical level categories derived to find phrasal categories defined over these. The mechanisms for finding lexical categories can be implemented as a neural network, which learns to classify words into syntactically interesting classes.

One feature of the present version of this iterative procedure which is not attractive is that there is no mechanism for correcting inaccuracies in early categories, based on an oversimple model of the rules of the domain, even when a more elaborate model of these rules has been derived. For example, the initial bigram model does not allow for the possibility that there are some surface forms (for example, FIRE) which correspond to more than one underlying lexical representation, with a different lexical category (in this case, NOUN and VERB). This difficulty can be overcome by using the observed context of ocurrance of the ambiguous word to disambiguate it. This can be achieved, as we noted above, by using the analysis of the similarity between phrases.

We hope that the general approach to the bootstrapping problem that we have outlined can be applied to other domains, as well as learning linguistic categories, and other problems involving the analysis of sequential structure. For example, in learning the structure of a visual domain, simple statistics concerning neighbouring values in the image (either grey scale values, or values which are the output of some pre-processing) can be used as basis for constructing statistical models of visually interesting categories. There will, of course, be no easy general solution to the bootstrapping problem - after all, this would be tantemount to a general theory of the processes of cognitive development or scientific enquiry. However, we hope that we have shown that in specific contexts, it is possible to bootstrap successfully using statistical methods.

References

Chater, N. (1989) Learning to respond to Structures in Time. Research Inititiative in Pattern Recognition: Technical Report RIPRREP/1000/62/89. Malvern, UK.

Chater, N. & Conkey, P. (in submission) Finding Linguistic Structure with Recurrent Neural Networks. Submitted to ICANN 1992, Brighton, UK.

Cleeremans, A., Servan-Schrieber, D. & McClelland, J. L. (1989) Finite State Automata and Simple Recurrent Networks. *Neural Computation*, 1, 372-381.

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

Elman, J. L. (1991) Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning*.

Finch, S. & Chater, N. (1991) A Hybrid Approach to the Automatic Learning of Linguistic Categories. *AISB Quarterly*, 78, 16-24.

Gold, E. M. (1967) Language Identification in the Limit. *Information and Control* 16, 447-474.

Kohonen, T. (1982) Self Organised Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, 59-69.

Osherson, D., Stob, M. & Weinstein, S. (1986) *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists.* Cambridge, Mass: MIT Press.

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

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

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

Svartvik, J. & Quirk, R. (1980) *A Corpus of English Conversation.* Lund: LiberLaromedel Lund.