NEURODYNAMICS *and* PSYCHOLOGY

EDITED BY

M. Oaksford and G.D.A. Brown

Department of Psychology,
University College of North Wales
Bangor

ACADEMIC PRESS
Harcourt Brace & Company, Publishers
London Boston San Diego New York Sydney Tokyo

natural language data. These last two avenues have recently been explored by Finch and Chater (1991, 1992, this volume) with encouraging results.

Chapter 12

Learning Syntactic Categories: A Statistical Approach

Steven Finch and Nick Chater

12.1 The bootstrapping problem

The acquisition of language is remarkably swift and successful despite the exquisite complexity of what is acquired and the incomplete and errorful character of the data upon which acquisition is based. The problem is particularly difficult, since both the categories over which linguistic rules are defined, and the rules themselves must be found (if both of these must be learnt, rather than being prespecified). That is, the learner faces a "bootstrapping" problem (Finch & Chater, 1991, 1992): linguistic rules presuppose the linguistic categories in terms of which they are stated; and the validity of linguistic categories depends on whether or not they support perspicuous linguistic rules. Given this interdependence of rules and categories, it is not clear how acquisition can occur, except by searching the vast number of possible of categories/rules combinations at once.

The bootstrapping problem arises in the acquisition of all aspects of linguistic structure, whether phonological, syntactic or semantic. Indeed, similar problems arise in learning the structure of almost any new domain. For example, in learning an academic subject, say elementary physics, learners must somehow acquire both the relevant concepts and the correct rules of inference defined over those concepts. For example, learners must grasp the concepts of momentum, force and so on, as well as the rules for how these concepts can be manipulated and interrelated. 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. The same problem occurs in science where it is necessary to simultaneously develop 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 inquiry and in individual cognitive development.

The fact that language acquisition appears to be so rapid and successful in spite of these difficulties suggests that much information about the nature of language must be innate. This putative innate linguistic knowledge will specify features that all natural languages must share and can be thought of as a "universal grammar" (e.g., Chomsky, 1980). The suggestion that much of language is not learned at all downplays the magnitude of the learning problem that the child faces.

Even given a large innate body of linguistic knowledge, however, the problem of learning a language still involves solving a formidably difficult bootstrapping problem. Even if all human languages have the same underlying structure, apart from certain syntactic parameters, and differences of vocabulary, phonology and so on, the superficial differences between languages remain vast. So, even if linguistic *categories* are prespecified, the learner still has to assign these categories to parts of what may appear to be an almost arbitrarily varied speech stream. For example, even if the child knows innately that there are nouns, it still has to determine which sounds of the language correspond to nouns. And even if universal constraints on linguistic *rules* are prespecified, the particular rules appropriate for any specific language must still be determined from observed utterances. The child faces a bootstrapping problem because assigning categories to portions of the speech stream, and determining the aspects of the rules of grammar particular to a given language are of course profoundly interdependent. To make matters worse, this problem must be tackled at the phonological, syntactic and semantic levels. Even if a very strong nativist position is correct, the child must still possess powerful mechanisms for language learning.

We shall argue that, despite appearances, the bootstrapping problem can be addressed by finding linguistic categories (at least to a good approximation) without making assumptions about the linguistic rules defined over those categories. Once an approximate set of categories has been fixed, rule learning can begin, and a mutual refinement of rules and categories becomes possible. We shall concentrate on syntactic categories, because they appear to pose the most difficult learning problem.

12.2 How might syntactic categories be learnt?

In learning a language, the child has two sources of information available: language-extrinsic information, concerning the observed relationship between language and the world; and language-intrinsic information, concerning the relation of fragments of languages to each other. Both sources must be drawn upon extensively. After all, learning semantics necessarily involves associating language and the world, and learning syntax requires learning intricate structural relations within language itself.

A natural assumption is that language internal information is the relevant source of information for learning *syntactic* categories and, indeed, the model that we develop below is concerned exclusively with language internal information. However, language-extrinsic information may, in reality, be a very significant source of information for the child, since there are strong correlations between syntactic and semantic categories (and hence between language internal and language external information). We shall see below there is considerable semantic information in purely language internal statistics: semantically related words such as numbers or compass directions tend to have the same linguistic distributions. Hence, in principle, useful semantic information could be gleaned from purely non-semantic observations of the relation between bits of language. Equally, since syntactic categories are correlated with semantic categories (extremely crudely, nouns refer to objects, verbs to actions or relations, adjectives to properties and so on), it may be possible to extract information about syntactic categories from the relationship between language and world. While this kind of "semantic bootstrapping" may be a significant factor in syntax learning (Pinker, 1979), we shall concentrate here on the contribution that purely language intrinsic-information could have. In particular, we shall investigate whether distributional evidence alone can be used to provide good approximations to syntactic categories.

Ideally, any language learning model should be tested on natural language input like that to which children are exposed. If the model were adequate, as the corpus to which the learner is exposed increases (and, if necessary, the model is modified to capture collateral changes in the child's cognitive capacities, such as, for example, working memory limitations), performance should gradually converge on adult syntactic categories. However, there is not a sufficiently large, or a sufficiently continuous, body of child and caregiver data to test a learning system over the entire course of development. This gives rise to two rather distinct projects. The first project is to show that the end state, the standard syntactic categories, can be attained in principle. This can be

assessed by testing the model on any reasonably large natural language corpus. The second project is to show that the methods used to reach this end state can account for the developmental trajectory of category acquisition. The computational work reported in this paper is concerned exclusively with the first project. Work on the second project, using the CHILDES language database (MacWhinney, 1989) is currently in progress.

The approach that we advocate uses statistical methods to learn the syntactic categories of English words and phrases from noisy text. We shall present this work in three stages. First, we outline and apply a statistical method for learning approximations to syntactic categories of lexical items. Second, we extend this approach to find the syntactic categories of short phrases. Third, we consider how these methods can be realised in a neural network, and give some simulation results. We finish by relating our approach to other work on computational modelling of language acquisition, and suggesting possible future directions for research.

12.3 Learning the syntactic categories of single words

Our method for learning syntactic categories involves three stages: (1) measuring the distribution of each word; (2) comparing the distributions between pairs of words, and (3) clustering together words with similar distributions. We shall consider each of these in turn.

12.3.1 Stage 1: Measuring the distribution of each word.

In traditional linguistics, words and phrases are categorised into several standard syntactic categories: nouns, verbs, noun phrases, verb phrases, and so on. One justification for this taxonomy is afforded by a number of "distributional tests," which assume that words and phrases that 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 concerning grammaticality. In the present context such intuitions cannot, of course, be presupposed, but a modified "statistical replacement test" is a good starting point:

> "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 assigned 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, an extremely simple definition of the context of a word must be assumed. In related and much earlier work on a small (15,000 word) corpus of child speech, Kiss (1973) defined context purely in terms of the probability of each possible immediate successor word, and found some structure in the resulting linguistic categories. Rather than record only the immediate successors of a word, we collected statistics for the preceding two and following two words surrounding the "focal" word. 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 for a given focal word can therefore be thought of as a vector composed of four sets of 150 values, each value corresponding to the frequency with which one of the 150 most common words appears in a given context position (preceding word, following word, last word but one, next word but one).

12.3.2 Stage 2: Comparing the distributions between pairs of words.

Having obtained a vector representing the distribution of each word of interest, we must compare distributions to see which words are likely to have the same syntactic category. In the spirit of the statistical replacement test described above, we propose that any reasonable measure defined to elucidate linguistic distributional similarity should be insensitive to the absolute frequency of occurrence of any particular word. In other words, it should be dependent on the relative frequency with which it co-occurs with other words. That is, it should satisfy the "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 measure of similarity."

There are several candidate measures for vector similarity which give results in quite good agreement with standard linguistic intuitions. In the experiments that we report below, we use the Spearman Rank Correlation Coefficient between the vectors of frequencies of context words, which produced the most satisfactory results. Since Rank Correlation between two vectors is in the range [-1,1], we used an appropriate rescaling of values into the range [0,1].

12.3.3 Stage 3: Clustering.

The measure of distributional similarity compares pairs of words. To divide all words into categories, clusters of similarly distributed items must be found. We used the most standard hierarchical clustering algorithm introduced by Sokal and Sneath (1963) which has been widely applied throughout the biological and social sciences. By using the distributional similarity metric as the basis for a hierarchical cluster analysis, words with similar distributions are placed nearby in the hierarchy. Nodes in the resulting taxonomy should correspond closely to traditional syntactic categories.

This simple method is surprisingly successful in practice. 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 learning syntactic categories in real corpora. We used a 40,000,000 word corpus of items from the USENET newsgroups, stripped of headers, footers and the like. This corpus is extremely heterogeneous, including formal and informal text on an enormous variety of academic, recreational and other topics. It is a very noisy corpus, containing numerous typographical errors, ungrammatical sentences, and all manner of idiosyncratic stylistic quirks. Yet 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. The three examples below show the ten words most similarly distributed to "three", "I" and "south":

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

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

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

At this level, syntactic categories and semantic relatedness are both apparent — numbers, personal pronouns and compass directions are all closely

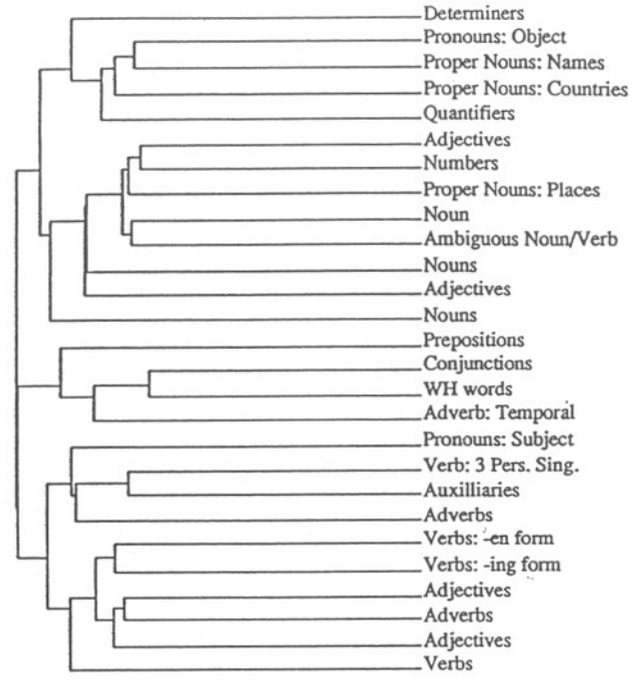

Figure 1. This is a summarised diagram of the clustered structure of 2000 words showing how interesting linguistic structure can be elucidated from such a structure. A small proportion (< 5%) of the data has either been omitted, or does not accord with the labels we use here.

associated. Notice, however, reasonably good distributional similarity does not necessarily imply strong semantic relatedness or even sameness of syntactic category — there appears to be no semantic relationship between "three" and "local," "I" and "didn't" or "south" and "school." A full cluster analysis is able to use the sum of pairwise associations between words to extract much better categories than these correlational data might suggest.

The cluster analysis produced a tree structure, or dendrogram, describing the relationship of the 1000 most common words in the corpus. This is, of course, much too large to display in a single diagram, so we first give an overview of the structure of the tree, before look at its fine detail. Figure 1 shows the large-scale structure of the tree — we have labelled each node

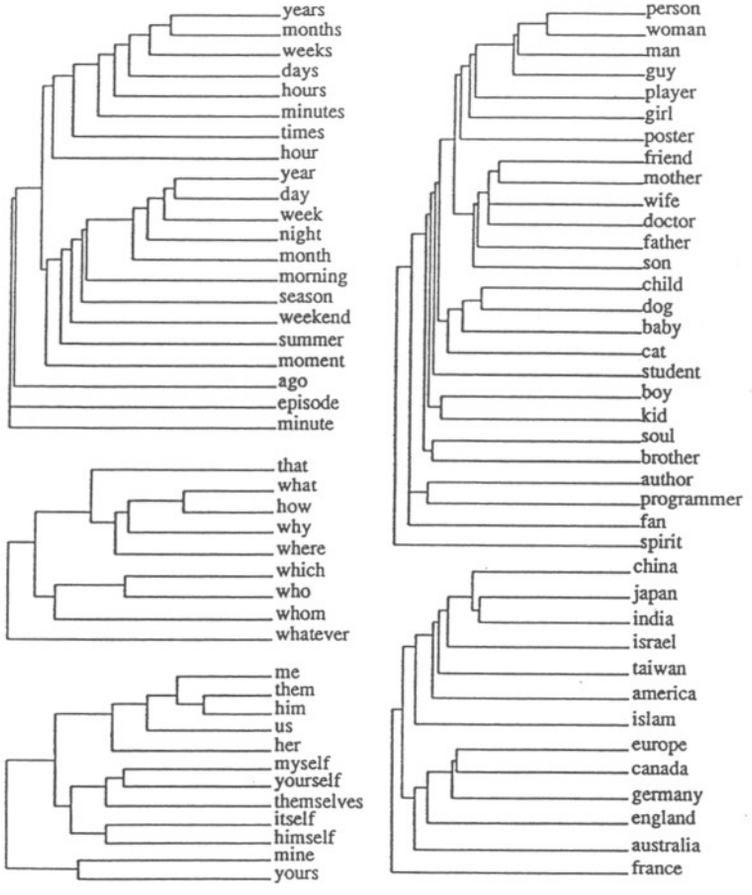

Figure 2. Dendrogram showing structure obtained via cluster analysis

according 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, fewer than 5% are misclassified with respect to the

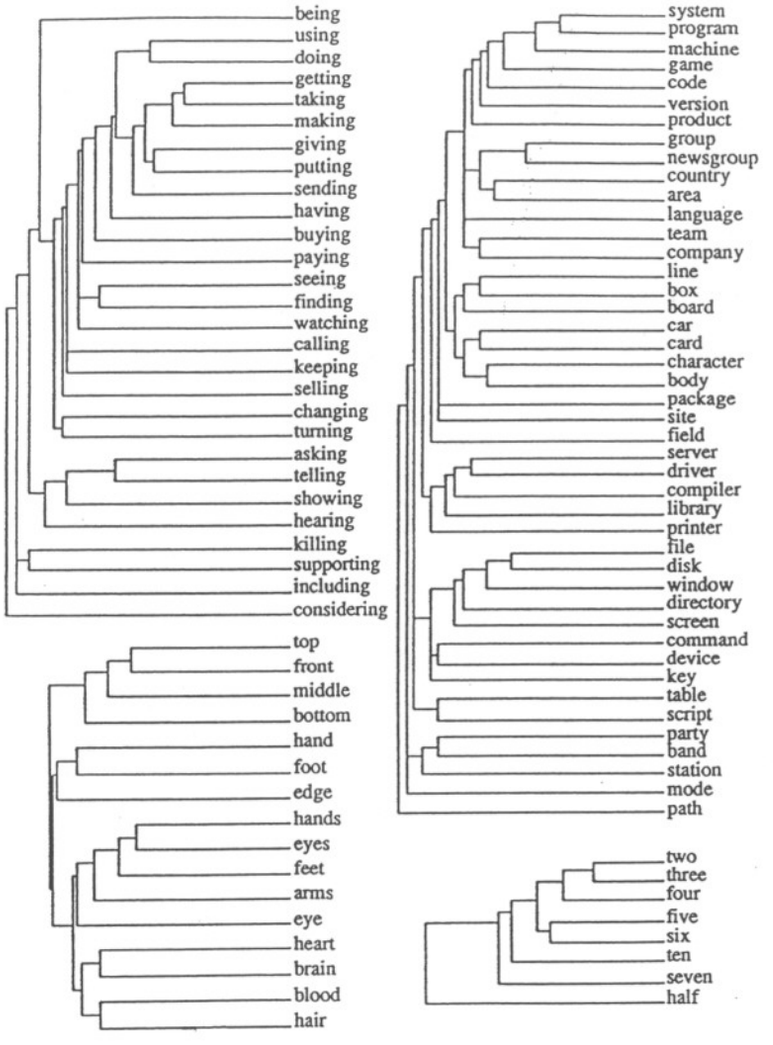

Figure 3. Dendrogram showing further structure obtained via cluster analysis.

label that we have given to their dominating node. Thus the gross taxonomy of the lexical items is very close to a standard taxonomy of syntactic categories.

Figures 2 and 3 show some of the low-level structure apparent within the dendrogram. These categories generally respect syntactic category, clustering together nouns, WH words, pronouns and reflexives, participles and so on. Perhaps more striking is the extent to which semantic factors are apparent. For example, periods of time, numbers, people or things treated like people (such as cats and dogs), and computer terms, are all grouped together, and are interestingly related. It is clear that there is considerable accord between empirical and syntactic/semantic similarity.

12.4 Learning the syntactic categories of phrases

So far, we have been concerned only with deriving a syntactic classification of single words. The distributional test in linguistics is, however, just as applicable to phrases as to single lexical items. It is therefore interesting to see whether noun phrases, verb phrases, adjective phrases and so on, group together when short sequences of words are classified on the basis of their distribution.

An immediate problem with observing the statistics of phrases is that since any individual phrase will occur very rarely, distributional statistics will tend to be extremely sparse and unreliable. To address this problem, we analysed not sequences of words but sequences of syntactic categories, derived from our previous analysis. We used 30 categories formed by cutting the dendrogram at a particular level of dissimilarity. In the original newsgroup corpus, individual words were replaced with a code that corresponded to the class to which they belong. For instance, each of the two-word sequences "the women," "the file" and "most data" was replaced by the sequence of labels "C30 C16." Unlike the sequences of words from which they are created, sequences of categories occurred frequently enough in the corpus to obtain reliable distributional statistics. In these experiments we classified sequences of between one and three words in length.

Presenting sequence data in dendrogram form becomes rather cumbersome, so instead we show some of the "tightest" clusters. That is, the dendrogram is "cut" at a particular level of dissimilarity and the sequences in that cluster are listed. 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 which occur as non-finite verbs and nouns with roughly equal frequency, 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 appears in the same contexts as non-finite verbs. Prepositional phrases and noun phrases also cluster together well:

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

It is possible to apply this procedure at a higher level still, using these short sequences as a starting point. We can cluster together short *sequences of sequences* of items, in terms of the distribution of short sequences in which they are found. This allows us to cluster together longer phrases (in the analysis that we have conducted so far, these phrases may be up to six words in length).

We briefly give some examples of the phrasal categories that this method can discover. First let us consider what we call *proto-sentences*. A proto-sentence is a phrase which could reasonably be thought to be a candidate sentence if

parsed out of context. For instance, the phrase *the man ate* would be a proto-sentence, even if it occurred in the context of *the man ate the apple* in which it would not be assigned the role of sentence. Also, because of ellipsis, NP movement, and the like, many sequences may be analysed as sentences which do not themselves stand alone as candidate sentences.

Here are some randomly chosen examples of proto-sentences from the category. We give examples in terms of phrases in the original corpus, again randomly chosen, which are members of the category, rather than in terms of sequences of syntactic labels:

what is a context
it might be a good idea
that's a different story
you see a problem
you will also receive a copy
that there was an error
the world isn't perfect
you start out
it really was lost
it does have a german title
we are looking
the government won't let them
it would be a good idea
i did notice it
i can get the book
you can actually see it
you need more information
this information is available
they were picked up
we could hold some events
you carry them
i just received my copy
you were found out
i just don't want it

Here is a random selection of sequences of words from a category whose tokens largely correspond to prepositional phrases:

out
out of this state
into a form
to those questions
of language and information
in the appropriate box
to a function
in school french
it out
by the way
of the terms
on its argument structure
of a variable
with this
of program performance
on the basis
of the file
in other words
to the development
in general
of such a news group
for this
on usenet
in areas of political rights
on the basis of religious law
up
to such a rule

The tokens here are almost exclusively either full prepositional phrases, or the first part of the prepositional phrase including the head noun of the rest of the prepositional phrase. The category of noun phrases includes the short noun phrases described above, and more complicated constructions such as Det NBAR PP, as in *the child of a woman* or *a piece of paper*, but not sequences such as *the man who I saw yesterday*, possibly because these are typically too long to be considered:

the reason
such questions
a moral law

the problem with it
a more accurate memory
the real number system
the article
it
many cases the option
a discussion on this
the child of a woman
a problem here
a gun
the six day war
his behavior during his life
his ideas about the rights
the four letter name
it for no reason
a piece of paper
someone at the post
some sources for your last statement

One limitation of the present version of the methods described so far is that each word is only assigned a single syntactic category (typically, its most frequent reading). Since many words have more than one syntactic category, it is important that the method can be augmented to capture other readings. One possible way to approach this problem is to use information concerning phrasal categories. For example, if an item or class of items occurs on both verb phrase and noun phrase type contexts, it may be appropriate to assume that it can function both as a verb and as a noun. On encountering a particular instance of the word, the appropriate reading could be chosen on the basis of the context in which it is found (although the sheer number of syntactically ambiguous words means that difficult combinatorial problems must be overcome). Whether this, or some other, method can successfully derive more all or most of the syntactic categories of a given word, rather than just the most frequent, is an important area for future work.

12.5 Neural network implementation

We have shown how it is possible to derive good approximations to the syntactic categories for English without having an account of the rules of syntax, by collecting statistics, deriving a similarity metric, and applying

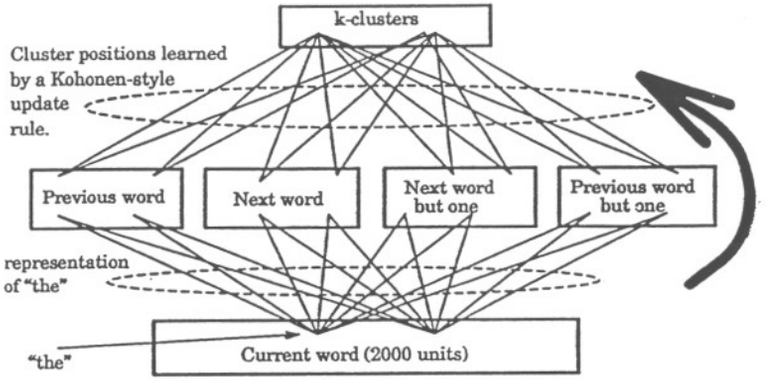

Figure 4. Network architecture used to implement the statistical algorithm.

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.

Providing a neural network implementation for these methods is an interesting challenge from a psychological point of view for at least three reasons. First, the existence of a neural network implementation demonstrates that the method can be parallelised using an array of very simple processing units. It this were not possible, and implementation required a very large number of serial steps, the method would be less plausible as a model of learning in the brain. This type of constraint is an example of how implementational considerations can influence higher level "cognitive" theory (Chater & Oaksford, 1990). Secondly, a mechanistic network model will give rise to processing and learning predictions which can be tested experimentally (for example, Seidenberg & McClelland, 1989). Thirdly, a concrete model can be damaged, and its characteristic patterns of performance breakdown observed, to assess whether the model can account for neuropsychological deficits (for example, Plaut & Shallice, 1993). Here we shall be concerned only with the first consideration, showing that a parallel implementation is possible, although further empirical validation of the model is a future concern.

Realising the statistical algorithm outlined above in a neural network requires the implementation of each of the three steps of the computation: collecting the distributional statistics, computing the similarity metric and performing a cluster analysis. The co-occurrence statistics are collected by a simple associative learning, using a completely localist representation of words. A bank of 2000 units represents the possible current words and four banks of the 150 units representing each of the context words, in each of the four context positions (Figure 4). This part of the network is trained by presenting successive focal words in the corpus with their appropriate contexts. Hence a single input unit is active at any given point, along with one unit in each of the four banks of context units. A Hebbian associative mechanism is used to train the weights between units representing current words. Hence the vector of weights from each input unit comes to represent the probability that each context word is present (in the appropriate location), given the current word. That is, the weight vector for each input unit comes to represent the distribution of contexts for the word for which that input unit stands. As usual with Hebbian learning, the weights are normalised, rather than being allowed to increase indefinitely. This factors out pure word frequency information, which we assume to be irrelevant to linguistic category, and thus satisfies the "replacement criterion" above.

Once learning is completed, the operation of the lower portion of the network is straightforward. A single input unit is turned on, representing a particular "current" word, and the linear units in the next layer are turned on in proportion to the frequency with which each context word has been paired with that current word in each of the four context positions. That is, the pattern of activation over the 600 units which form the output of this lower portion of the network represents the probability distribution of contexts associated with the current word. Each of the 2000 possible current words is associated with a distinct distribution of contexts, in such a way that words which appear in similar contexts (and hence are likely to be of the same syntactic category) are given similar patterns of activation.

The next stage in the process is to compute a measure of the similarity of the distributions associated with each word. Rather than use rank correlation, we use the similarity metric implicit in the neural network representation and hence no computational work is required. We have experimented with this metric in the statistical analysis reported above, and found that although it is successful for simple artificial data sets, it is inferior to rank correlation for real natural language data. Hence the clustering achieved by the network should

match less well with standard linguistic categories than does the clustering obtained using purely statistical methods.

The final stage in implementing the statistical analysis in a network is to carry out a cluster analysis on the words on the basis of distributional similarity. The upper part of the network uses a simple unsupervised clustering method 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. Notice that this style of clustering partitions the words into a distinct categories, rather than providing a full hierarchical analysis.

Before performing large-scale simulations, we tested a small network on the task of clustering together letters by distributional similarity. When the network consisted of two output nodes, and hence was forced to find two clusters of letters, it precisely divided vowels from consonants. Again with two output nodes, we clustered together phonemes rather than letters, taking data from a small (12,000 phoneme) corpus of phonemically transcribed speech (taken from Svartvik & Quirk, 1980). Here, too, the network 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, the full-scale network and corpus were used. The first layer of the net was trained with the 40,000,000 word newsgroup corpus. After training, when a "current word" is presented, the middle layer represents the distribution of contexts in which that word occurs. The patterns representing the distribution of each of the 2000 words under consideration were then clustered into 100 groups using a Kohonen network.

We found that 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. Furthermore, some clusters appear to correspond to no linguistically coherent category. Some examples of the

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

clusters formed are shown below and it is clear that they reflect both syntactic and semantic similarity:

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 bring

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

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 version 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 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

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 we saw in Figure 1). Hence it may be possible to improve network performance further.

This approach to building a neural network model exemplifies a general methodological strategy that may often be useful. Just as with designing symbolic computer systems, we have found it valuable to study *what* computation must be performed to solve the problem in hand, before turning to consider *how* that computation can be implemented (although, of course, the answer to the *what* question is sought with an eye to whether or not a satisfactory *how* implementation will be possible). Specifically, it has been useful to consider the statistical problem that must be solved to learn syntactic categories from distributional information, and then to choose a network implementation which embodies an appropriate statistical method.

A neural network which discovers phrasal categories could be constructed in the same way as for lexical level data, by implementing the steps in the statistical analysis in a neural network. This is possible using approximately the same architecture as before, and the resulting clusters would be composed of short sequences of words rather than single words. Such a network has not, however, been implemented.

12.6 Relations to other approaches

The work reported in this paper is based on the assumption that language acquisition involves solving difficult statistical problems. In this section, we shall clarify what the statistical approach to the acquisition involves, and relate our work to other work on the computational problem of acquiring language from language data.

The view that language acquisition involves solving difficult statistical problems is easily confused with two very different and much more contentious claims: the empiricist claim that no language-specific innate information is used in child language acquisition, and the view that language itself should be modelled statistically. These various claims can, and should, be kept separate, however. As we have stressed above, whether or not the child has access to innate information, the problem of using a noisy corpus of language data to pick out a specific natural language is still a formidible statistical problem. Furthermore, the statistical perspective on the form of the language acquisition *problem* does not dictate that the child's *solution* involves applying standard methods in mathematical statistics. All that is important is that the statistical problem is solved, and the language learned from the data available, by whatever means. Nonetheless, it is worth exploring whether standard statistical techniques which have been developed to solve other problems in which underlying models must be uncovered from noisy data may provide insights into aspects of the problem of language acquisition, and we have followed this approach in the work reported above.

Turning to the second point, that language *acquisition* is statistical does not entail that natural language should itself be described in purely statistical terms. Thirty-five years of work on generative grammar have given ample reason to assume that many important aspects of natural language are best described in terms of, and are likely to be generated and understood by using, complex systems of rules. In particular, language structure and language processing cannot be understood in terms of simple stochastic mechanisms such as Markov (1913) sources. What is statistical is not the model of the language itself, but the problem of finding a grammar for language, given a set of linguistic data.

With these considerations in mind, we now relate our work on learning syntactic categories to other work applying statistics to natural language, to neural network approaches and to formal language learning theory.

12.6.1 Statistical approaches to language learning

While, as we noted above, there is no necessary connection between viewing language acquisition as statistical, and viewing language itself as statistical, many applications of statistical techniques to natural language do model language as a stochastic process (Garside, Leech & Sampson, 1987; Jelinek, Lafferty & Mercer, 1990; Markov, 1913; Shannon, 1948, 1951). Typically, language is assumed to have been generated by a simple parameterised model. Accordingly, learning involves adjusting the parameters to fit the observed data as well as possible. Common models include hidden Markov models (Huang, Ariki & Jack, 1990) and stochastic context-free grammars (Booth, 1969). In the domain of syntax, the main justification for assuming such unrealistic models is simply that they are a mathematically and computationally well understood starting point for research.

It has recently become computationally feasible to apply such models to large text corpora. For example, Garside *et al.* (1987) learn how to disambiguate parts of speech using the bigram statistics of local context from a tagged corpus (i.e., a corpus with correctly disambiguated parts of speech); Brown *et al.* (1988) show how it is possible to translate, to some extent, between English and French using a Shannon-style noisy channel model, and training on a large bilingual corpus. Jelinek *et al.* (1990) have trained stochastic context-free grammars using large corpora of English.

While much of this work has aimed simply to improve performance in particular computational application domains, these techniques have more recently been applied actually to learn linguistic structure (Brill, Magerman, Marcus & Santorini, 1990; Church, 1988; Pereira & Schabes, 1992). Of particular relevance in the present context is the recent application of statistical methods to find linguistic categories from untagged natural language corpora (e.g., Kneser & Ney, 1991; Marcus, 1991). What is distinctive about our work is that our motivation is to find categories which accord with an appropriate linguistic taxonomy, rather than finding categories which are useful from some practical point of view. Nonetheless, the techniques used in practical computational linguistic tasks may be of relevance to more psychologically and linguistically motivated work.

12.6.2 Neural network approaches

What amounts to an alternative, rather non-standard, statistical approach to language learning has been developed within the neural network tradition. Here the underlying model of the language (or part of the language) to be

learnt is a complex non-linear system of equations, which correspond to a system of simple processing units connected in parallel by real valued weights.

The most influential approach to learning the structure of sequential language-like material, and, in particular, the categories which reveal that structure, is due to Elman (1990, 1991; see also Chater, 1989; Cleeremans *et al.*, 1989), using his simple recurrent neural network (SRN) architecture. The SRN is typically trained to predict the next element in a sequence of inputs generated by a simple grammar. It can develop patterns of hidden unit values which, when appropriately averaged and cluster-analysed, reveal underlying syntactic categories.

Another approach to learning linguistic categories uses a competitive network to produce to topographic mapping between the distribution of contexts in which an item occurs and a two-dimensional space (Ritter & Kohonen, 1989, 1990; Scholtes, 1991a,b). The results show that items with the same linguistic category tend to lie in neighbouring regions of the space, although there is no algorithm for finding an actual linguistic classification from this data.

These neural network methods are performing particular sorts of statistical analysis on artificial language data. For a small-scale case, it has been shown that the categories implicit in the hidden unit values of the SRN reflect certain distributional statistics of the training data (Chater & Conkey, this volume). Furthermore, topographic mapping methods also appear to have a statistical interpretation — as a non-standard method of multidimensional scaling on the basis of distributional similarity.

Both approaches face two important difficulties. First, it has not yet been possible to scale up from very small artificial data sets to deal with real linguistic data. For example, in SRNs, which rely on prediction, learning becomes extremely inefficient and slow, if it occurs at all, as the language become more complex, and prediction becomes more difficult (Chater & Conkey, this volume). Secondly, the linguistic categories are implicit within the network, and can only be revealed using a subsequent cluster analysis. Thus, a significant amount of the computational work in finding syntactic categories is not performed by the network itself. Both of these limitations are overcome by our "direct" neural network implementation of a proven statistical algorithm — real natural language corpora can be used, and the network itself classifies words into syntactically interesting classes.

12.6.3 Relation to formal language learning

In this section, we shall compare the statistical approach to language learning with the formal language learning tradition, which is inspired not by probability theory and statistics but by automata theory and logic (e.g., Osherson, Stob & Weinstein, 1986; Pinker, 1979, 1984). Both approaches are at a level of abstraction which is very far from dealing with the details of child language acquisition, yet both attempt to inform that study. We argue that the standard idealisation of the problem of language learning within formal language learning theory abstracts away not just aspects of the language learning problem that make the problem more difficult, but also aspects that make language learning easier. By taking into account the statistical structure of the language learner's input, the nature of the learning problem appears to be rather different, and more tractable. This does not mean, of course, that the important limitative results derived within the formal language learning tradition can be disregarded with impunity, or that ideas from this tradition cannot be fruitfully applied to the idealisation of the language learning problem that the statistical approach assumes. We suggest that it is important to complement and enrich formal language learning theory with statistical ideas, rather than to replace it wholesale.

Formal language learning theory has focused on the acquisition of a generative grammar from a set of sentences (and sometimes explicitly marked non-sentences) of the language. Sentences are chosen from possible sentences of the language, usually with equal probability, with restrictions ensuring that every sentence of the language must be chosen eventually, and so on.

Importantly, this input is error-free, and does not contain analogues of the ungrammatical, inchoate or unfinished utterances which are typical of natural language. If formal language learning accounts are extended to include the possibility of error in the input data, then the learning mechanisms which are generally considered break down. In particular, even if the correct model of the language is somehow found, it will be rejected as soon as an ungrammatical sentence (for which it is, rightly, unable to account) is encountered. To avoid this difficulty, what is required is some mechanism for distinguishing between signal (the grammatical sentences of the language) and noise (the ungrammatical non-sentences) and to assess which grammar best accounts for the available data. In short, a statistical approach to language learning is required.

There is, however, a different reason to employ statistical ideas. By removing the statistical structure present in real natural language a great deal of information which may be useful in learning is lost, and learning may thus

be made more difficult. We shall consider just one example of how abstracting away from statistical information makes learning harder, concerning the role of negative evidence in language acquisition.

An important early result (Gold, 1967) was that no infinite language (that is, no language containing more than a finite set of sentences) can be learnt from "positive" evidence alone, that is, simply from examples of sentences in the language. The problem stems from the fact that overgeneral grammars cannot be disconfirmed by positive evidence alone. There is no way that the learner can rule out sentences that are not grammatical simply because they do not appeared in the corpus — for since the language is infinite, there will be infinitely many grammatical sentences of the language which also do not appear in any finite corpus. If negative evidence is allowed, according to formal language learning accounts, language learning becomes much easier. For example, there are computable methods for learning classes of languages with an infinite number of sentences, such as finite state and context-free languages. Unrestricted transformational grammars are still not learnable, even with negative evidence, although when such grammars are constrained, learning may be possible.

These considerations lead to the expectation that negative evidence plays a crucial role in children's acquisition of language. The obvious way in which evidence could be provided would be in adults giving differential feedback to grammatical and ungrammatical sentences uttered by the child. However, the empirical evidence does not appear to support the view that adults make such a differentiation. Brown and Hanlon (1970) analysed parent–child interactions and found no correlation between parental approval and grammaticality of the child's utterance, or between the appropriateness of adults' answers to a child's questions and the grammaticality of the question. Other studies have confirmed these results and show that even when parents are sensitive to the grammaticality of their child's utterances, the resulting feedback to the child is extremely variable from one occasion to the next, dependent on the child's age, from child to child and so on (Demetras, Post & Snow, 1986; Hirsh-Pasek, Treiman & Schneiderman, 1984). It appears therefore that negative evidence does not play an important role in language acquisition. This finding has been taken to back up the claim that a great deal of linguistic knowledge must be innate, since without such knowledge, and with only positive evidence to take into account, learning even the simplest language is impossible.

If the statistics of natural language data are available, however, negative evidence is not necessary to disconfirm grammars of the language which overgenerate. For example, the learner can disconfirm the overly general

hypothesis that all possible strings of words are syntactically legitimate simply by noting that this hypothesis does not explain why it is that some sentences occur with high frequency, and some do not occur at all.

An important contrast between the formal language learning and the statistical approach is that the former has no measure of *how well* the observed sentences of a corpus fit with the grammar that is currently under consideration: a corpus fits a grammar if and only if all the sentences that it contains can be generated by that grammar. This is why a grammar which massively overgenerates with respect to a corpus (for example, which generates all possible sequences of words) can never be disconfirmed on positive evidence alone.

Statistical methods provide ways of measuring how well a grammar fits a corpus, as a special case of measuring how well a model fits a body of data. For example, from the point of view of Bayesian statistics, this measure is particularly simple — it is the probability of the grammar given the corpus. This can be computed, by Bayes's theorem, from appropriate assumptions about the prior probability of the grammar and the probability of the corpus given the grammar. This latter quantity will automatically be small for overgenerating grammars. Since they allow a large set of sentences, the probability that the particular set of sentences in the corpus will be obtained is reduced. It is for this reason that Bayesian statistics is said to embody Occam's razor, automatically punishing overgeneral models (MacKay, 1991; Skilling, 1989). This means that overgenerating grammars can be rejected (albeit provisionally) on statistical grounds using positive evidence alone.

While the use of statistical information in place of negative evidence is attractive in principle, its practical application is very difficult because the space of possible grammars is so vast. Thus strong innate constraints on the nature of language may be still be necessary to explain how language is acquired. Furthermore, a full Bayesian analysis of the probability of even a single grammar is computationally very expensive and could at best only be approximated in real language acquisition. Despite these problems, it should be remembered that the problem of language acquisition is immense from any perspective — this very fact should persuade us that no potential source of information to the learner should be ignored.

12.7 Future directions

We have shown that it is possible to find good approximations to linguistic categories from raw text using statistical information, and sketched how such

methods can be realised as neural networks. In doing so, we have suggested how it is may be possible to avoid the bootstrapping problem.

The problem of learning the grammar of natural language given the syntactic categories is, while easier than before, still enormously difficult. Suppose, contrary to fact, that our category learning mechanism could correctly assign syntactic categories to words and phrases of all lengths, and appropriately resolve syntactic ambiguities (needless to say, it is entirely unclear whether this level of performance could be achieved using extensions of the methods that we use, or using just distributional information at all). If this were possible, the learner could derive something approximating to a linguistically motivated tree structure for each sentence encountered. That is, learning could proceed from a *tagged* corpus, rather than a raw stream of words. Even were this possible, the problem of language learning would still be extraordinarily difficult.

One way to proceed would be to feed the tagged corpus into a symbolic algorithm which constructs a phrase structure grammar for the language. Each cluster is associated with a symbol X of the grammar, which may be rewritten as any element of the cluster, perhaps the string $Y_1, \ldots, Y_n$. Since this set of rules reflects only the broad statistical regularities in language, this resulting grammar will tend to overgeneralise very strongly. This would give a set of rules of the form $X \rightarrow Y_1, \ldots, Y_n$. It would then be possible to "fine-tune" this grammar by assigning probabilities to each rule and adjusting these parameters using the standard training algorithm for a stochastic context-free grammar (Booth, 1969). Were such an approach to prove feasible, and it is certainly a long way off at present, it would usefully exploit the category learning in deriving linguistic rules. There are no existing statistical algorithms which are able to learn stochastic context-free grammars from raw data, since the search problem entailed by having neither categories nor rules established is so vast. But the grammar that could result from this kind of analysis will be extremely simplistic compared to the grammars postulated by modern linguistics. There are great numbers of subtle and important linguistic regularities that a stochastic context-free phrase structure grammar is unable to capture — but for more realistic grammatical formalisms, no learning algorithms exist, even for tagged corpora.

Even if the bootstrapping problem can be solved, and syntactic categories learnt without making assumptions about linguistic rules, the problem of grammar learning remains intractable. This is not, however, an adverse comment on the power of statistical methods. The problem of language learning is a statistical problem of perhaps unparalleled complexity. To unravel

the methods that must be involved in solving it will stretch current statistical methods to their limits and beyond.