A Hybrid Approach to the Automatic Learning of Linguistic Categories

Steven Finch*& Nick Chater†

October 27, 1991

Abstract

Symbolic and neural network architectures differ with respect to the representations they naturally handle. Typically, symbolic systems use trees, DAGs, lists and so on, whereas networks typically use high dimensional vector spaces. Network learning methods may therefore appear to be inappropriate in domains, such as natural language, which are naturally modelled using symbolic methods. One reaction is to argue that network methods are able to *implicitly* capture this symbolic structure, thus obviating the need for explicit symbolic representation. However, we argue that the *explicit* representation of symbolic structure is an important goal, and can be learned using a hybrid approach, in which statistical structure extracted by a network is transformed into a symbolic representation. We apply this approach at several levels of linguistic structure, using as input unlabelled orthographic, phonological and word-level strings.

1 Introduction

Since the Chomskian revolution it has been recognised that the structure of human language is enormously complex. This realisation, while stimulating for the development of formal symbolic models of syntax, has been taken to have strong nativist consequences for models of language acquisition (Gold 1967; Pinker 1984; Osherson, Stob & Weinstein 1986). Overgeneralising somewhat, experience has proved that modelling language successfully requires the use of complex, structured symbolic representations such as trees, DAGs or lists; and that current non-nativist approaches to learning are sensitive only to statistical regularities. For example, the statistical methods of learning used by neural networks are widely considered to founder on the problem of learning structured material (Lachter & Bever 1988; Pinker & Prince 1988). It might therefore be concluded that simple statistical methods which may be unable to capture the full richness of the symbolic structure of natural language can have nothing interesting to say about how that structure is learned.

The goal of this paper is to provide an illustrative example of how very simple statistics, which can be easily computed by a network architecture, can play a key role in learning structural properties of natural language. For example, we present a hybrid model which uses a neural network to manipulate certain statistics which are then translated into symbolic form. This method is based on collecting and clustering bigram statistics using a rank correlational metric. In principle, then, it is sensitive only to very local dependencies. Nonetheless this very impoverished informational measure is, in practice, sufficient to discover the difference between verbs, adverbs, adjectives,

*University of Edinburgh, Centre for Cognitive Science.
†University of Edinburgh, Department of Psychology.

prepositions, determiners, and between singular and plural nouns. In fact, this statistical method is also able to reveal apparently semantic properties of words, grouping numbers, compass direction and animate objects together. It is important to recall that the model derives these categories without any prior knowledge of the syntax of or the syntactic categories of the data concerned.

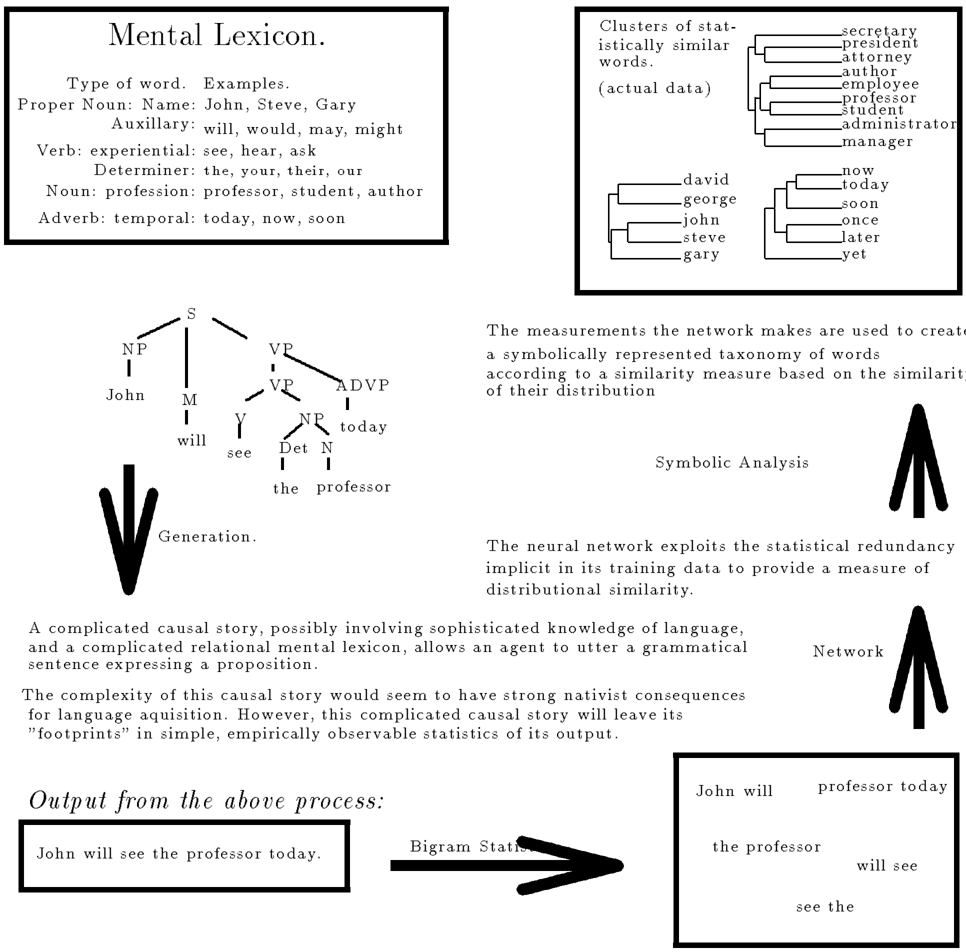


Figure 1 shows our conception of how this work on learning relates to conventional models of language production. Where the left hand side of figure 1 denotes the *generation* of language using conventional rules and representations, the right hand side denotes the *analysis* of that language by network methods combined with a method of converting the value of certain statistics into a structured representation. The generated sentence reflects the underlying structure of the rules and representations that gave rise to it only very indirectly. One reason that statistical or network methods are appealing in this context is that the performance data is so untidy that a non-statistical rule based approach is liable to reject correct hypotheses concerning aspects of language structure in the face of apparently contradictory data, which is actually simply the caused by performance errors. The simple bigram statistics that we use (detailed below) are denoted in the bottom right hand corner of the figure. Notice that all information about the overall order of the string of words has now been irretrievably lost; all that is known is how often each pair of words were observed to be in a certain relationship (in this case "next word"). These observations are a bigram statistic of the language, and aspects of the structure of the lexicon (in this case, syntactic category) is recovered by first computing a correlational measure using a neural network, and then transforming the result into a tree structure, which provides a structured taxonomy of the lexicon.

The structure of the paper is as follows: in the next section we discuss the general nature of the problem of learning linguistic categories from raw data, and discuss a previous approach to the same problem, due to Elman (1990). In section 3. we outline the hybrid method that we used and sketch how they can be implemented in a neural network architecture. In section 4. we present the results of some computational experiments with a variety of linguistic data. In the conclusions, we suggest that this hybrid strategy might be extended to other areas where structured representations might be learned from statistics computed by a neural network.

2 The problem and a previous approach

Models of language learning have typically been concerned to learn rules of syntax, given syntactic categories, and have not been directed at the problem of *finding* the syntactic categories of natural language from raw word level data. We focus on this important and complimentary aspect of language learning, which is a prerequisite of learning the syntax of the language and assume that the system has no prior knowledge of the nature or number of syntactic categories.

Within the neural network tradition, the most influential approach to the computational problem of learning linguistic categories is due to Elman (1990) who used a recurrent neural network and cluster analysis to reveal syntactic and semantically based clusters of "lexical" items.

Elman generated a continuous stream of input using a grammar of two and three word sentences with a vocabulary of X items, and a very small number of rules, which encoded certain semantic dependencies. A recurrent network was trained to predict the next element in this continuous sequence. Although it was impossible to perform this prediction perfectly, as the sequence is non-deterministic, the network nonetheless was able to exploit some of the predictability in the sequence, and developed interesting hidden unit representations in doing so. Elman analysed these representations by calculating the mean state of the hidden units when each particular item of vocabulary was presented as the current input, and hierarchically cluster-analysing the resulting vectors of values. This cluster analysis grouped together syntactically and semantically similar words, which was taken to show that the network implicitly encodes the categories underlying this simple language. Further studies have also used the same technique applied to material with a somewhat more complex syntax (Elman 1989) and using finite state grammars (Cleeremans, Servan-Schreiber & McClelland 1989).

The fact that the hidden unit representation *implicitly* encodes the syntactic/semantic structure is not, however, necessarily evidence that the recurrent network has learnt to extract this structure. Experiments reported below show that raw unlabelled input data also implicitly encodes such information. Using this observation, we developed a more direct approach to finding linguistic categories, which elucidates the nature of the processing used, and leads to a system which can be applied to real corpora in many linguistic domains.

Rather than cluster analyse a complex and little understood measure, such as the mean value of the hidden units of a recurrent network, it may be more profitable to use better understood statistical methods like cluster analysis on the input data itself. One of the simplest partial descriptions of the structure of any sequence of symbols are its bigram statistics. We used cluster analysis on bigram statistics of a large of corpus of informal written material to attempt to derive various linguistic categories, including syntactic categories for English. This approach can be realised in a hybrid system using a network to decide how similarly distributed items are and applying a standard clustering technique to the output of the network. While Elman sees cluster analysis simply as a means of analysing network behaviour, we see it as an integral part of our model, since our goal is to derive an explicit symbolic taxonomy of words which captures linguistic regularities in real data sets. Cluster analysis plays the role of the interface between a vector space representation in the network and a structured symbolic representation of the data.

3 A hybrid system

Rather than using a standard network model and using statistical techniques to analyse its behaviour, we take our motivation from standard statistical approaches to sequential material. We derive a statistical method for finding linguistic categories and show how an important component of this method can be realised in neural network hardware.

3.1 Statistical motivation

One standard test in theoretical linguistics for words and phrases having the same syntactic category is that they are similarly distributed. That is, if it is possible to replace one phrase with a particular string of words in any syntactic environment, then this string has the same syntactic category as the phrase. (see, e.g., Radford 1988). Distribution is usually thought of with reference to linguistic *possibilility*, rather than about distributions observed in actual data. By constrast, our approach interprets distribution in terms of statistics of actual corpora.

Our approach thus departs from the standard notion of distribution in two ways. Firstly, theoretical linguists have not been concerned with actual frequency of occurrence of words or phrases in particular contexts, but rather with which sequences of words are judged to be grammatical. By contrast, we focus precisely on the statistics of observed errorful linguistic data. We are interested in how this untidy performance data can be used to cast light on the underlying linguistic competence that gives rise to it. Secondly, theoretical linguistics has usually been concerned to distinguish syntactic reasons why a string is not possible from semantic/pragmatic reasons (Chomsky 1957) - due to the assumption of the autonomy of syntax. For instance, if “mat” and “idea” are assigned the same syntactic category then if “the cat sat on the mat” is syntactically acceptable, then “the cat sat on the idea” must also be syntactically acceptable. This sentence is however anomalous and linguists would standardly rule it out for semantic reasons, such as selection restrictions (Katz & Fodor 1963) or meaning postulates (Dowty, Wall & Peters 1981). Our method does not distinguish between syntactic and semantic/pragmatic factors in determining our measure of similarity of distribution. Nonetheless, as we shall show below, syntactic factors are sufficiently predominant to allow the recovery of syntactic categories. Thirdly, where theoretical linguists are typically concerned with the distributional characteristics of *phrases*, which may consist of many words, our analysis, at least in the first instance, works at a word level, and makes no assumptions about the phrasal structure of the language.

An easily computed, well-understood and theoretically neutral statistic of natural language is given by what are known as “N-gram” statistics. Roughly speaking, we shall treat two words as having similar linguistic distributions if their “N-gram” statistics are sufficiently well correlated. The “N-gram” statistics can be collected, the correlations can be calculated by a simple neural network, and cluster analysis applied to its output.

3.1.1 N-gram statistics

An N-gram is an ordered sequence of N symbols (words, letters or whatever). The frequencies of occurrence of each N-gram in a continuous stream of data constitutes the N-gram statistics of the data set. The 1-gram statistics of a data set are therefore simply the frequency of each symbol in the data set. If the data set is natural language, 1-gram statistics amount simply to a table of word frequencies. The 2-gram or “bigram” statistics of a data set are the observed frequencies with which each pair of words appear consecutively or with a fixed number of intervening words.[1]

[1] In this paper, we shall be interested in bigram statistics where the words are adjacent or separated by a single word. The notion of a bigram is considerably more general than this, and the relationship between pairs of words can be determined in a wide variety of ways.

So, for example, in the sentence "To be or not to be", the bigram statistics for the relation "immediate successor" are: [to be]:2, [be or]:1, [or not]:1, [not to]:1 and 0 for all other possibilities. In the analysis below we shall also measure the frequencies of pairs of words which have a single intervening word.

Notice that these bigram statistics ignore all structure in the data set which is not determined by dependencies between adjacent pairs of words, or pairs of words separated by only one word. For natural language, in which dependencies may exist between words which are sensitive to arbitrarily much intervening context (Chomsky 1957), this structure will be very important for a complete analysis.

Bigram statistics can be collected by a neural network in which units in one layer represent the "current" word, and units in the other layer represent the values of the previous, next, last but one, and next but one words. This means that the contingency table can be interpreted as a weight matrix, which is updated by a fixed amount whenever both the units to which it is connected are on (see Figure 2). That is, the network uses a Hebbian learning rule.

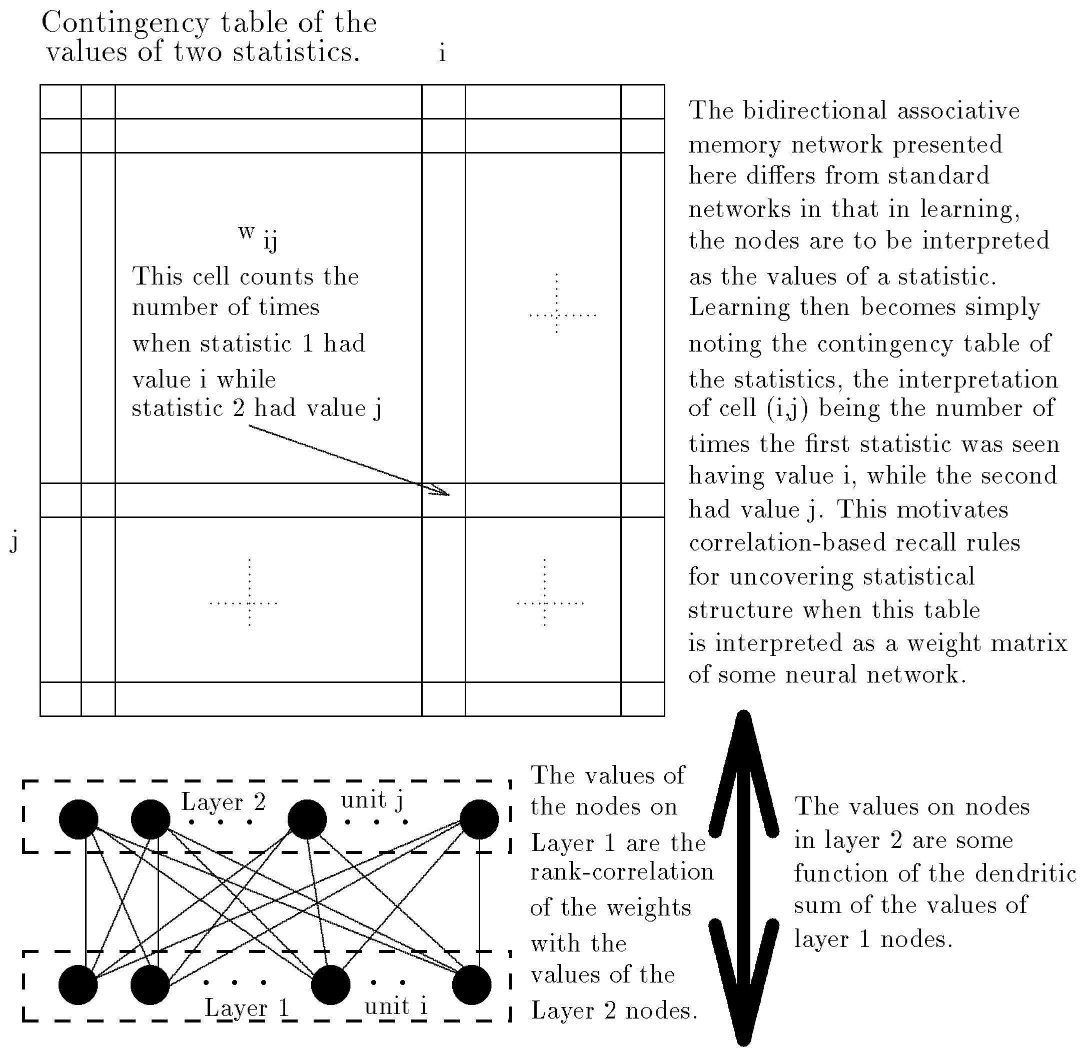


Figure 2

The measure of similarity of distribution that the network calculates can be interpreted as a statistical measure of correlation between the bigram statistics of each word. The measure which is found to be best empirically (that is, which revealed categories most in accordance with linguistic theory) is a standard non-parametric measure, the Spearman Rank Correlation Coefficient. From the statistical literature, we know that on data which is known not to be well modelled by a multivariate normal distribution, tests based on rank often more powerfully distinguish between statistical hypotheses (Hettmansperger 1986), and this is evidence that correlation measures based on rank may be better at uncovering underlying structure than other measures.

This correlation is calculated by nodes in layer 1 of the network (see Figure 2) calculating not a function of the dendritic sum, as in most networks (see for example Hertz, Krogh & Palmer 1991), but instead finding the correlation between values of units in layer 2 and the weight vector it has learned. In operation, a unit in layer 1 of the network (which corresponds to a particluar lexical item) is activated and excites units in layer 2 according to the probability distribution of values of the following/preceeding/next-but-one/last-but-one words. This layer in turn reactivates layer 1, since connections are reciprocal. The activity of a unit in layer 1 is therefore a measure of the correlation of its distribution with that of the initially activated word. Thus, after one iteration, the network has derived a measure of distributional similarity for a particular word, which can then be used by the symbolic cluster analysis component. Alternatively, activation can be allowed to iterate back and forth through the network several times, with the effect of bringing similar items more closely together. We shall return to the utility of using several iterations below.

3.2 Symbolic cluster analysis

A standard way of presenting a finite number of data points in high dimensional vector space is in the form of a tree or *dendrogram* (Sokal & Sneath 1963). This is sometimes known as hierarchical cluster analysis. The tree is generated so that it represents similar items, according to the chosen metric, as nearby leaves in the tree, in the sense that they the leaves share a nearby common branch. The distance along the horizontal axes of the dendrograms that we shall present give a quantitative measure of similarity.

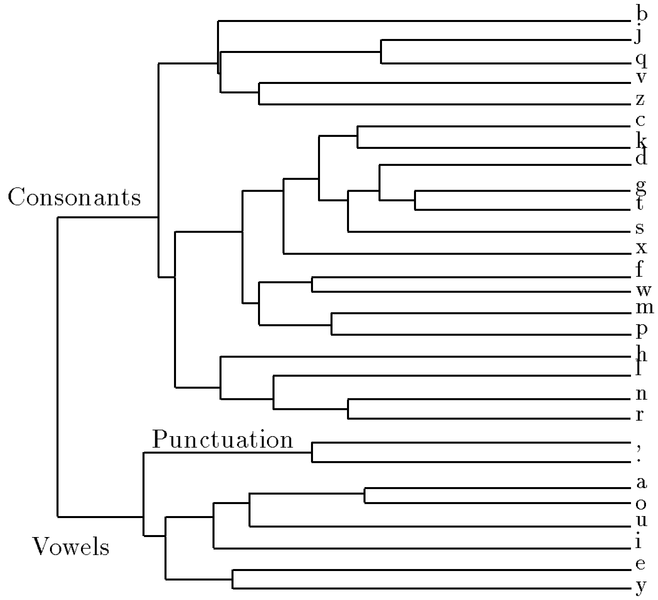
The results of cluster analysis are sensitive to the metric of similarity used. We shall see below that the correlational metric used by a single iteration of the network gives a less tightly clustered tree than if the network has been allowed to iterate several times. This indicates that the iteration of the network is playing a useful computational role.

4 Computational experiments

Rather than give a detailed account of the computational experiments that we have performed, we give a brief summary of some of principle results obtained so far.

4.1 Analysing letter and phoneme data

As a first experiment, we analysed distribution of letters rather than words, using a 1,300,000 letter corpus taken from Usenet newsgroups. Figure 3 shows the dendrogram obtained. The first (and hence most fundamental) division in the tree is between consonants and others. On the "others" side of this division, the next principal division is between punctuation marks ("." and ",") and vowels ("a", "e", "i", "o", "u", "y").



We also performed a similar analysis on a small corpus of phonemically transcribed speech, collected from a variety of speakers, taken from the Lund corpus (Svartvik & Quirk) the output of which is shown in Figure 4. The corpus is sufficiently small that some phonemes are only represented by a few occurences. Nonetheless, again broad distinctions emerge between vowels and consonants. Although one phoneme is spuriously classified as a vowel, it only occured eight times in the entire corpus of 15,000 phonemes.

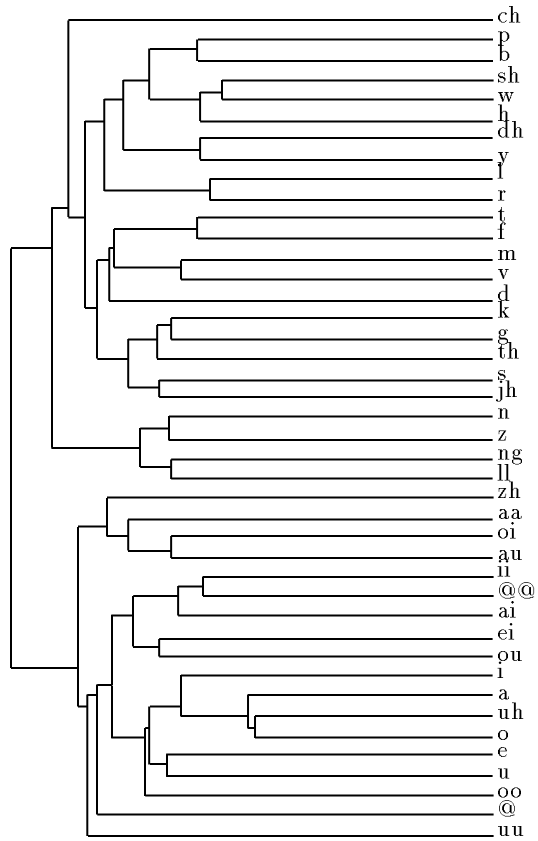


Figure 4

4.2 Syntactic categories in natural language

We took a 33,000,000 word corpus of Usenet newsgroup data, stripped of headers, footers and the like. Even before cluster analysis, a list of the ten nearest neighbours of sample words shows that the 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.

These lists appear to show effects of both syntactic and semantic similarity. The first reveals a high correlation between "number" words. Numbers higher than six are not frequent enough to be considered in the data set, and "one" has a rather different distribution, since it has more than one grammatical function, (a matter which we shall consider further in the conclusions). That "three" does not correlate well with "two" is more surprising. The second list shows the distributional similarity of pronouns, and the third shows that very fine-grained semantic information (i.e., being a compass direction) can be detected.

4.3 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. We therefore an overview of the structure of the tree, 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 examples, single letters of the alphabet, 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 5 therefore shows the gross taxonomy of the lexical items for the newsgroup corpus. Thus, this taxonomy is very close to a standard linguistic conception of the different species of syntactic category and how closely they are related.

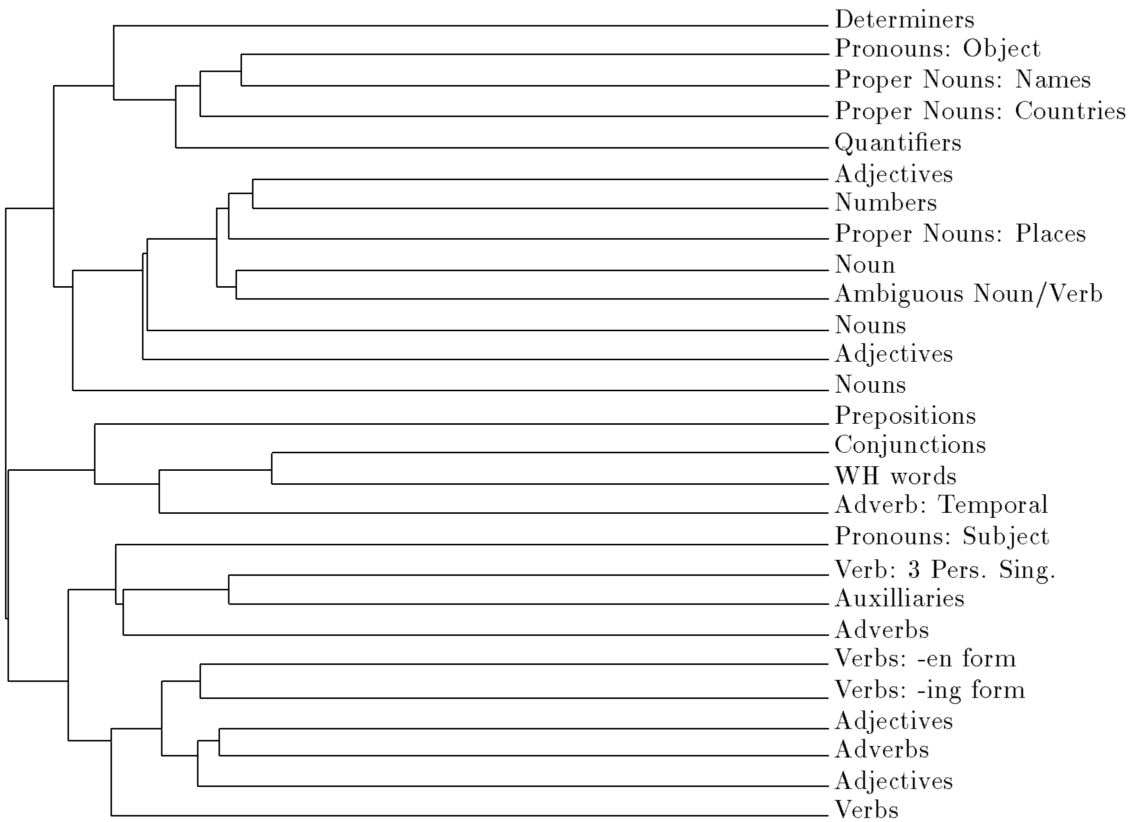


Figure 5

This is a summarised diagram of the clustered structure of 1000 words showing how interesting linguistic structure can be elucidated from such a structure. A small proportion (¡ 5%) of the data has either been ommited, or does not accord with the labels we use here.

Figure 6 shows the structure of two subtrees within the whole tree. The left hand tree shows the clustering within the category of prepositions. It is interesting to note that words which are semantically closely related appear very close together in the tree (e.g., up/down, inside/outside) although this is not so in all cases. The right hand tree shows a subcluster of nouns, the plural nouns. General semantic regularities are apparent — plurals which denote people or computer terms are grouped together. Also, pairs of items which are semantically related again tend to appear very close in the tree (e.g., women/men, articles/postings/comments, states/countries).

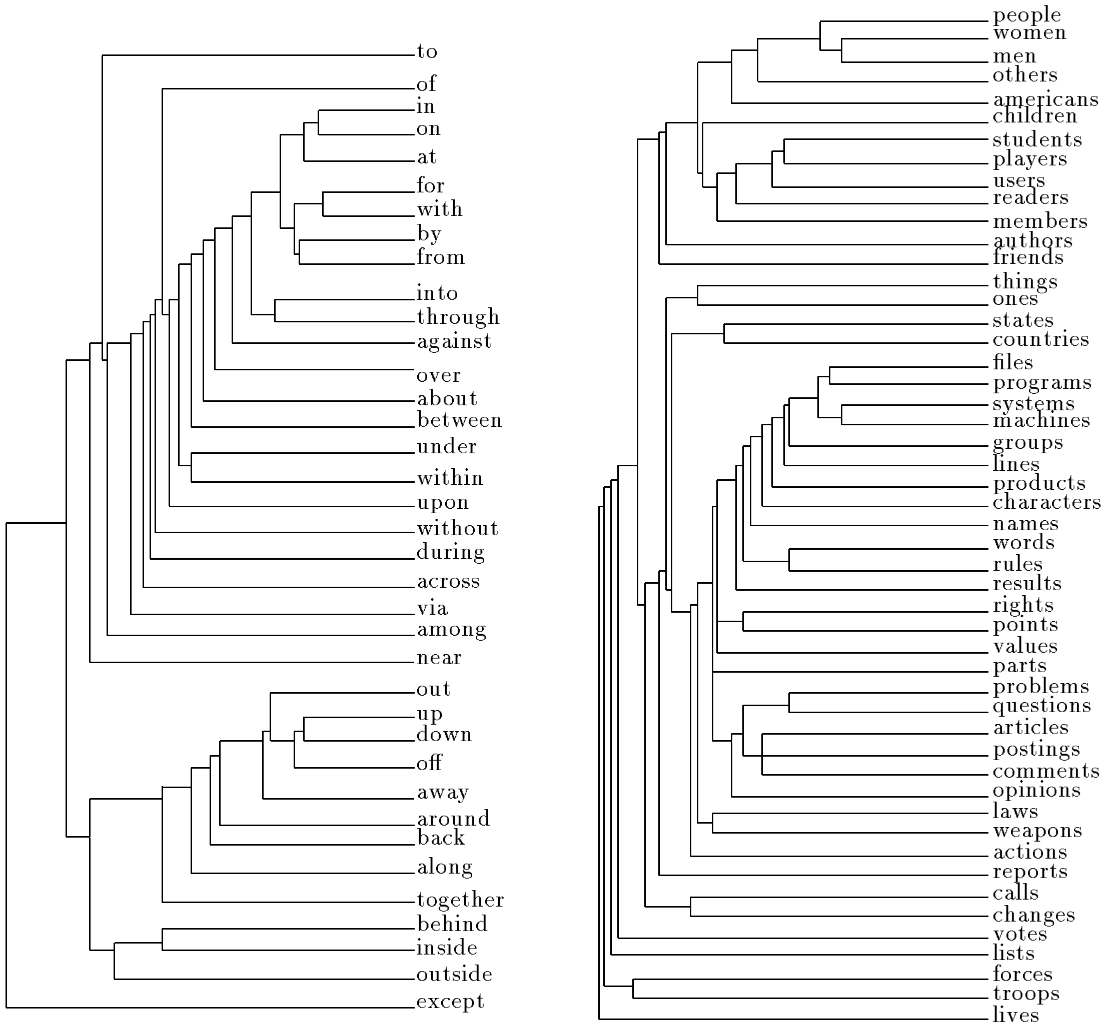


Figure 6

The left hand part of Figure 6 shows the structure within the preposition node of Figure 5. The right hand part of the figure shows the subtree of plural nouns from within the noun node in Figure 5.

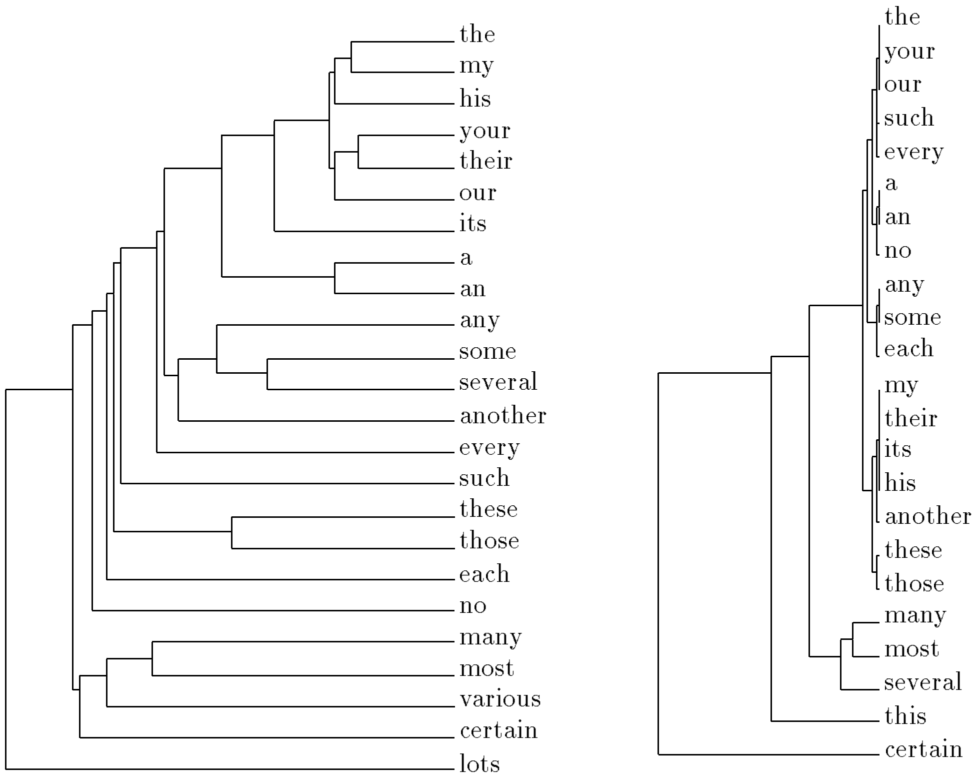


Figure 7

Figure 7 shows a comparison between the cluster of determiners after 1 and 10 iterations through the network. The network can therefore be thought of as transforming the structure of the data (in this case, "flattening" the structure so that similar items are made much more similar).

These figures are produced by cluster analysis of the output of the network after a single iteration. If the network is allowed further iterations, it tends to group linguistically similar items more closely together. This is shown in Figure 7, which shows the clustering of determiners after 1 iteration and after 10 iterations.

5 Future directions

There a number of ways in which this work can be naturally extended. An obvious extension is to see to what extent such methods can be applied to corpora of data from languages other than English. A particularly interesting question is to what extent word order data can reveal syntactic and semantic features even in language such Finnish and Latin where lexical (morphological) structure plays a more important role than word order in determining syntactic structure. Further, rather than being limitted to the particular bigram statistics that we used, we might consider many other statistics, and see which can be exploited to produce interesting clusters. Also informational measures over classes of statistics can be exploited to identify "optimal" statistics.

More generally, instead of considering the distributions of words, we might consider the distributions of sequences of words. This might seem to be infeasible since the frequency of any given sequence will be low, even in a very large corpora. This difficulty can be circumvented by using the results of analysis at the word level to tag the corpus into a relatively small number of categories, and to consider sequences of these categories. The expectation is that cluster analysis of such sequences may reveal higher level linguistic categories such as noun-phrase, sentence and so on. This process can itself be performed iteratively, where sequences of only 1, 2 or 3 items need at first be considered, since, in conventional linguistics at least, phrasal constituents typically decompose into 2 or at most 3 sub-constituents. Preliminary results are encouraging, although it is not yet clear to what extent it will be possible to build up something approximating to a phrase structure grammar for the data set.

Apart from the use of N-gram statistics, the methods used here are not specific to sequential data. Perhaps it will be possible to find statistics of other domains which can be exploited in a similar manner to uncover underlying symbolic structure. It seems likely that the best way to pursue such an approach will be to employ a hybrid approach: use a neural network to find statistical regularities in the data, and then derive a symbolic structured representation from the network's output.

References

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

Chomsky, N. (1957) *Syntactic Structures.*, The Hague: Mouton.

Dowty, D. R., Wall, R. E. & Peters, S. (1981) *Introduction to Montague Semantics.*, Dordrecht: Reidel.

Elman, J. L. (1989) Structured Representations and Connectionist Models. *Proceedings of the Cognitive Science Society of America* 17-23.

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

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

Hertz, J. A., Krogh, A. & Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation.* Redwood City, Calif: Addison-Wesley.

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

Katz, J. J. & Fodor, J. A. (1963) Structure of a Semantic Theory. *Language*, 39, 170-210.

Lachter, J. & Bever, T. G. (1988) The Relation between Linguistic Structure and Associative Theories of Language Learning: A Constructive Critique of some Connectionist Learning Models. *Cognition*, 28, 73-193.

Lehmann, E. L. (1975) *Nonparametrics: Statistical Methods Based on Ranks.*, San Francisco: Holden-Day.

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.

Pinker, S. & Prince, A. (1988) On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition, *Cognition*, 28, 195-247.

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.