Distributional Information and the Acquisition of Linguistic Categories: A Statistical Approach

Martin Redington[1]
Department of Psychology
University of Edinburgh
7, George Square
Edinburgh, U.K.
EH8 9JZ
frankred@cogsci.ed.ac.uk

Nick Chater
Department of Psychology
University of Edinburgh
7, George Square
Edinburgh, U.K.
EH8 9JZ
nicholas@cogsci.ed.ac.uk

Steven Finch
Centre for Cognitive Science
University of Edinburgh
1–4, Buccleugh Place
Edinburgh, U.K.
EH8 9LW
steve@cogsci.ed.ac.uk

Abstract

Distributional information, in the form of simple, locally computed statistics of an input corpus, provides a potential means of establishing initial syntactic categories (noun, verb, *etc.*). Finch and Chater (1991, 1992) clustered words hierarchically, according to the distribution of local contexts in which they appeared in large, written English corpora, obtaining clusters that corresponded well with the standard syntactic categories. Here, a stronger demonstration of their method is provided, using 'real' data, that to which children are exposed during category acquisition, taken from the CHILDES corpus. For 2·5 million words of adult speech, clustering on syntactic and semantic bases was observed, with a high degree of clear differentiation between syntactic categories. For child data, some noun and verb clusters emerged, with some evidence of other categories, but the data set was too small for reliable trends to emerge. Some initial results investigating the possibility of classifying novel words using only the immediate context of a single instance are also presented. These results demonstrate that statistical information may play an important role in the processes of early language acquisition.

Acquiring linguistic categories

In acquiring language the child solves an enormously difficult problem: To uncover an exquisitely complex model of language structure from a corpus of observed language which is both extremely partial and riddled with false starts, slips of the tongue, and ungrammatical sentences. One small, but crucially important, aspect of this problem is that of acquiring linguistic categories. Unless linguistic categories are established, learning how those categories may be composed together to uncover the linguistic rules of the language appears to be impossible.

[1]Experimental work was performed by MR, supervised by NC and SF, and supported by SERC studentship No. 91425111, and ESRC studentship No. R00429234268.

The difficulty of the problem of language acquisition is often taken to suggest that the child must be able to draw upon an extremely rich store of innate information or Universal Grammar (UG) (Chomsky, 1980). According to strongly nativist views, the very phrase "language learning" embodies a misconception; it is assumed that the most complex and intricate aspects of language are not learned but innately specified. In the case of the early stages of the acquisition of linguistic categories, however, it seems that prior information can be of little value. Suppose, for example, that the existence of a particular stock of syntactic categories (noun, verb and so on) is known *a priori*. There still remains the difficult problem of deciding which words of the language have which syntactic categories, since vocabulary is clearly not innately specified. Furthermore, the problem of assigning words to categories does not appear to be appreciably simplified by the prior knowledge of the stock of linguistic categories and their role in UG. This is because the universal grammatical features of language can only be mapped on to the specific surface appearance of a particular natural language once the identification of words with syntactic categories has been made. Of course, once some identifications have been successfully made, it may be possible to use prior grammatical knowledge to facilitate further identifications. To solve the problem of establishing initial linguistic categories, however, it seems that the contribution of innate knowledge must be relatively slight (although see Lasnik (1989) for some tentative suggestions of how innate constraints might contribute).

It seems, therefore, that the problem of learning linguistic categories may involve the use of unsupervised learning methods which embody little prior information about the structure of the language. We shall describe computational experiments which show that distributional information present at the word level in English speech can be exploited in order to reveal the underlying syntactic categories, although the methods we discuss are applicable both at other levels, and to other types of linguistic category.

Language Internal and Language External Cues

We distinguish between two sources of information that might be used in order to learn syntactic categories: Language internal—dependent on the relationships between elements (in this case, words) of linguistic input; and language external—concerned with the relationships that can be uncovered between language and the world.

Here, we consider only language internal information. Whilst language external factors are presumably of considerable importance, they are very difficult to model computationally, given our almost complete lack of knowledge as to how they can be appropriately represented. Empirical data concerning the child's representation of the world remains both anecdotal in nature, and difficult to interpret.

The unsupervised analysis of language internal relationships, in the absence of a priori information, falls naturally into the domain of statistical, or distributional, analysis.

Previous Computational Approaches

The statistical analysis of natural language. While the use of neural network methods has largely been restricted to data generated by artificial grammars, it has been possible to apply statistical models to large text corpora. Much of this work has aimed simply to improve performance in particular computational application domains, but more recently, these techniques have been applied to actually uncover linguistic structure (*e.g.* Brill *et al.*, 1990): That is, the goal is to automatically generate linguistically justified categories and rules. 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.* Marcus, 1991; Kneser & Ney, 1991). The algorithm described below (Finch & Chater, 1991, 1992, 1993) derives what is probably the most linguistically motivated taxonomy of categories to date.

Relevant neural network research. There has also been considerable interest in learning syntactic categories from scratch with the neural networks literature. While early neural network language processing models have been hard-wired (*e.g.* Hanson & Kegl, 1987), it has recently become possible to use learning methods to extract linguistic structure from (artificial) data. In particular, there has been significant interest in the finding the syntactic categories of lexical items.

The most influential approach is due to Elman (1990), using his simple recurrent 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 activation which, when appropriately averaged and cluster-analysed reveal underlying syntactic categories. A *post hoc* analysis of the statistics that the network picks up shows that the hidden unit patterns reflect a simple distributional property of the corpus; the vector of conditional probabilities of each possible next item in the sequence, given the preceding sequence of items (Chater & Conkey, 1992). Thus a simple statistical representation of each word by its conditional probability mirrors the hidden unit structure very closely.

This analysis feeds naturally into the present work. When attempting to learn a language whose structure is unknown, it is not of course possible to compute the conditional probability of each word given arbitrary context; however, it is possible to compute these conditional probabilities directly, based on local context, and this distributional statistic provides the basis of our statistical method outlined below.

Elman's approach is limited in that it fails to scale up from small, artificial data sets, to deal with real linguistic data. SRNs, rely on prediction, which rapidly becomes extremely difficult, since learning is inefficient and slow, if it occurs at all. Other neural network approaches (*e.g.* Scholtes, 1991) share similar limitations.

The Present Approach

A standard test in theoretical linguistics for words sharing the same syntactic category is similarity of distribution: If all occurrences of word A can be replaced by word B, without loss of syntactic well-formedness, then they share the same syntactic category (*e.g.* Radford, 1988) . The method here was to form a representation of the distribution of contexts within which each word appeared, and then to identify categories as corresponding to clusters within the space of possible distributions of context. The local context—the two preceding and two succeeding words was used, thus encompassing the short phrase structure constituents, within which dependencies between words are relatively highly constrained.

In practice, the most frequent 1,000 words in the target corpus were chosen as the focus words, and the most frequent 150 were chosen as the context words. The dependencies between the focus words and the context words were recorded by incrementing the value of the cell indexed by the appropriate focus and context word, in a contingency table corresponding to the appropriate context position; last but one word, previous word, next word, or next but one word. Thus for each focus word, the row of each contingency table forms a 150-dimensional vector, representing the observed distribution of each of of the 150 context words in that position. Stringing the vectors from each table together results in a 600-dimensional vector, representing the distribution of local contexts within which each focus word

appeared[2]. Given these context vectors, a number of clustering algorithms can be used to identify clusters of words occurring in the space of possible distributions of context, according to some chosen distance or similarity metric. Spearman's rank correlation coefficient, a robust similarity metric, has been found to give good (linguistically appropriate) results with natural language corpora, with clustering performed by hierarchical cluster analysis.

Using this method, Finch & Chater (1991, 1992) obtained excellent results using 30–40 million word corpora taken from Usenet newsgroups, with dendrograms produced by hierarchical cluster analysis corresponding well with a standard syntactic taxonomy; at a high level, nouns, verbs, adjectives *etc.*, were all differentiated. At a lower level, semantic regularities were apparent, *e.g.* computer-related nouns, countries, compass directions *etc.*, were clustered together.

One advantage of the method is that the context vectors can be easily constructed in a single pass through the input corpus, with space and computational requirements increasing linearly with the total number of focus and context words. This allows its use with large natural language corpora. The method is not limited to the word level—Finch & Chater (1991) demonstrated its ability to distinguish between consonants, vowels, and punctuation at the letter level, and between consonants and vowels in phonemically transcribed speech. In Finch & Chater (1992), again using a Usenet corpus, one, two, and three word combinations of syntactic categories (identified by a similar analysis at the word level) were classified according to their phrasal category (noun phrase, verb phrase *etc.*).

A weakness of the method is the absence of any adequate quantitative measure of the "goodness" of the resultant clustering (although the results so far are qualitatively impressive). The syntactic ambiguity of many words, *e.g.* "*Fire* the gun", "light the *fire*", renders quantitative assessment problematic. Generally the method will cluster such words with the category in which they most frequently occur in the input corpus, with, for instance, a cluster of ambiguous noun/verbs forming for words where neither form dominates. Additionally there is the problem of appropriately partitioning the space of possible distributions of context into disjoint categories, and the question of whether this is an appropriate aim, given the occurrence of syntactic ambiguity. However, it seems likely that the (good) first approximation to categories provided by the method may be sufficient for other, more sophisticated techniques to proceed from thereon.

[2]This can be implemented as a neural network (Finch & Chater, 1992), in which nodes in the first layer represent the current word, and those in layer 2 represent the previous but one, previous, next, and next but one words, with the bidirectional weights between the nodes being incremented appropriately, via a Hebbian learning rule. Thus by activating the node representing a word in layer 1, the activation vector of layer 2 is equivalent to the 600-dimensional vector representing the distribution of contexts. The layer 2 activation vector can then serve as the input to a Kohonen type clustering algorithm.

Results with the CHILDES Corpus

Whilst the method has been shown to be effective with written English corpora, the language to which children are exposed in early life is considerably different. Child-directed speech in particular is known to possess distinctive characteristics; in vocabulary and grammar it is a subset of the full adult system, and sentences are relatively short, simple, grammatical, and repetetive (Kuczaj, 1982). The success of the method with data of this type provides a stronger demonstration of its validity.

The results described here were obtained with data taken from the CHILDES corpora (MacWhinney and Snow, 1985). CHILDES is a machine-readable collection of corpora of child and child-related speech, transcribed by a number of investigators. The English language transcriptions involving non-impaired speakers were prepared, each utterance being indexed by age and sex of speaker, this information being taken from the documentation accompanying the transcriptions. The resultant corpus contained over 4·3 million words of speech, from nearly 6,000 speakers. The only preprocessing that was performed was to strip away the CHILDES coding information and punctuation present, and thus the corpus was rather noisy; for instance, 'mummy' and 'mommy' were effectively different words, as were 'dats' and 'thats', *etc.*

A number of analyses were performed on subsets of this corpus, the best results being obtained from the analysis of the adult speech only. Whilst there is no guarantee that the whole of the adult speech in the adult corpus was child-directed, it would seem to form a fair representation of the speech to which a language-learner might be exposed. The adult speech corpus comprised over 2·5 million words, from 3,416 speakers, 1,022 male, 2,141 female, and 253 whose sex had not been recorded.

The high level structure of the dendrogram formed from the adult speech is illustrated in figure 1. The dendrogram has been cut at a level of dissimilarity chosen to reveal the 'best' syntactic grouping. Of the 42 disjoint clusters thus obtained, the 13 shown contained over 90% of the focus words, and the labels for each of these clusters were appropriate for around 95% of the words in each cluster. The clusters containing 2 categories in the top half of the tree (*e.g.* pronouns/auxiliary verbs) do separate appropriately within the cluster. However such clusters are small in size compared to the noun and verb clusters, which contained over 70% of the focus

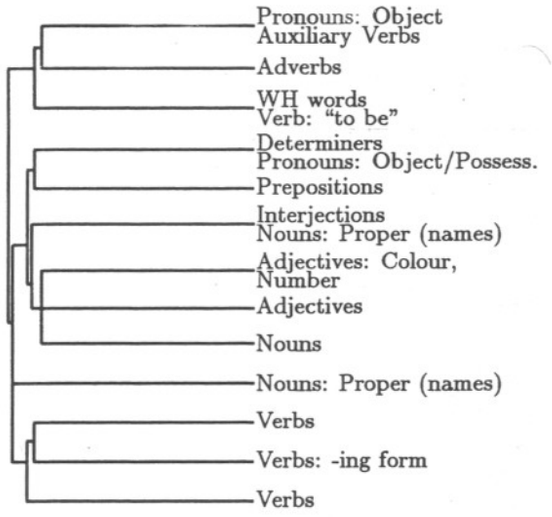

Figure 1: A summary diagram of the high level structure of the dendrogram for the adult speech. Approximately 15% of the data has been omitted, or does not accord with the labels used here.

words, and are highly coherent, containing the lowest proportion of inappropriate words. Subclusters within the noun and verb clusters show a distinction between singular and plural nouns, and, to some extent, between present and past verb tense, and the '-ing' verb form.

Figures 2, 3 and 4 illustrate a number of subclusters of those indicated in figure 1. The goodness of the clustering in figures 2 and 3 is typical of that throughout the noun and verb clusters. These results compare favourably with those of Finch and Chater (1991), especially given the much reduced sample size (2·5 million words, versus 30 million words).

It is not appropriate to attach too much importance to the general overall structure of the dendrogram, or to the position of any particular word within it—the former may change according to the particular set of words that are included in the analysis, and the latter may be affected by an idiosyncratic usage by even one individual; for instance many of the names that occur in the set of focus words are identifiable as arising from specific individuals. What is important is the general high level of correspondence between clusters and syntactic relationships.

Analyses were also performed on corpora consisting of the child speech, broken down by year of age. The results of these were much less clear-cut. The size of the samples was very small (2,000 words for children of a year or less, 88,000 for 1–2 year's, and approximately 0·5 million for each of 2–3, 3–4, and 4–5 year olds). As the method is dependent on the size of the sample, in order to gain a representative distribution of contexts for each word, this variation in corpus size no doubt affected the results. However, for the children of 2 years and older, noun and verb clusters were evident, although much less clear cut than for the adult speech. There was also some tendency for other categories to be evident—for instance, the 2-3 year old's dendrogram clusters together 'down', 'up', 'out', 'off', 'away', 'back', and 'home', and this tendency appears to increase with age. However, in the absence of a good quantative measure of the 'goodness' of clustering this has not been confirmed statistically. Analysis of the entire corpora yielded results of slightly lesser quality than those for the adult speech alone, unsurprisingly given the presence of the much noisier and syntactically less well-formed child speech.

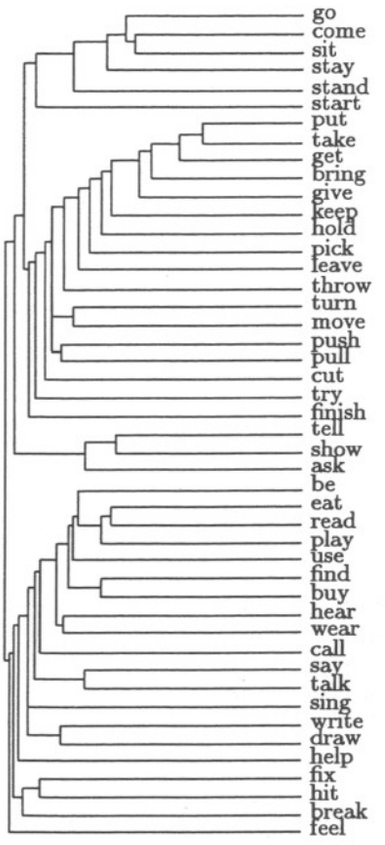

Figure 2: A typical subtree of the verb branch of the dendrogram. To some extent semantic clustering is evident (*e.g.* push/pull, say/talk/sing), and it occurs to this degree throughout the dendrogram.

Categorising New Words

Once the child has acquired the initial categories, there remains the problem of assigning categories to novel words, such as in the Berko test, where, given a single sentence such as 'John *glinned* the dog', the learner is required to recognise '*glinned*' as a verb and to use it appropriately (Maratsos, 1988). We have conducted a few preliminary experiments with

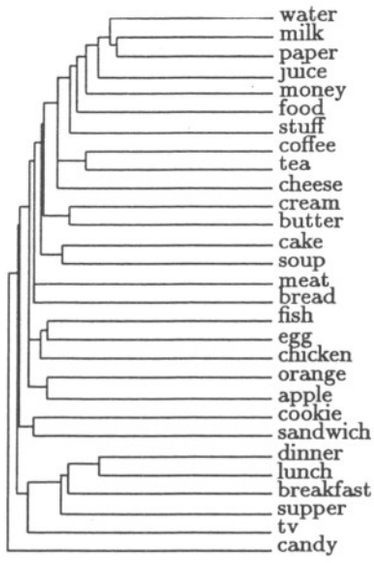

Figure 3: This noun branch subtree illustrates the capturing of semantic as well syntactic regularities. This is probably the best example of semantic clustering from the adult data, although the low frequency of non-nouns is typical of the noun cluster.

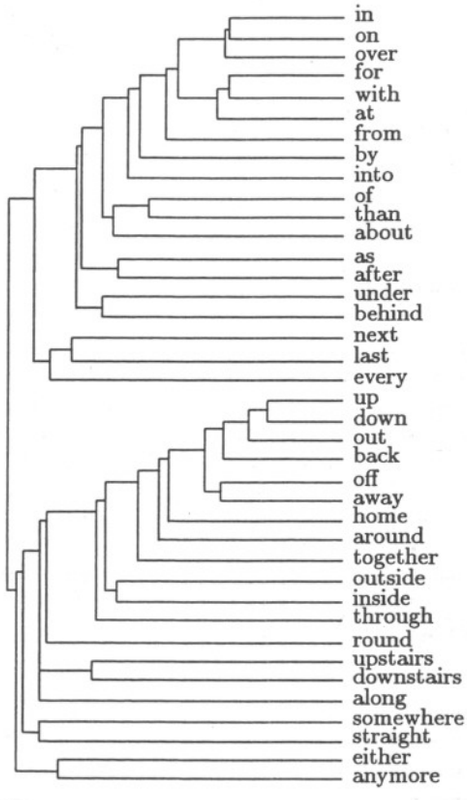

Figure 4: This is the entire 'preposition' branch, as indicated in figure 1.

CHILDES data, in order to see if slight variant of Finch and Chater's method can be used to achieve the recognition component of this task.

The method used was as follows; a set of 100 target words was selected, on the basis of frequency (between 70 and 77 occurrences in the adult speech), none of them being members of the set of focus words. A context vector was constructed for each of the occurrences of each target word, in the adult speech. Being based on a single instance, these 600-dimensional vectors were extremely sparse, with a maximum of four non-zero elements, where all four surrounding words were members of the set of context words (which was the same as that for the adult speech above). Thus the process so far was exactly as was performed for the entire adult corpus, except that the 'corpora' here were 5 words long, *e.g.* 'orange it tastes very sour' was an single occurrence, and the associated context for the target word 'tastes'.

Each occurrence's vector was then compared to the context vector for each of the original focus words, and the dot product of the two vectors calculated. This can be seen as a measure of the probability that each focus word would be found in the same context as the occurrence of the target word. The target word was then 'recognised' as belonging to the category of the focus word for whom this value was highest. A separate categorisation was also assigned on the basis of the three nearest neighbours—if two or more of the 3 'closest' focus words belonged to the same category, then that category was assigned to the target word. These categorisations were then compared to the categories to which the target words were assigned by hand. For the 25% of target words that were not members of the noun or verb categories, there was little evidence that they could be distinguished by this method. The percentage of occurrences correctly classified for nouns and verbs are shown in table 1. Also shown are the percentages of occurrences of each category that were misclassified as nouns, instead of as verbs, and vice versa.

Whilst these results are not impressive in comparison with other automatic classification techniques, given the paucity of the information utilised, the

Table 1: Percentage of occurrences classified as nouns or verbs, by 'actual' class (in capitals). Note that the VERB row total for 3 nearest neighbours exceeds 100%: One of the categories was 'ambiguous noun/verb', and neighbours in this category counted as both—thus some words were classified as both nouns and verbs.

Method:	by closest word		3 nearest neighbrs.	
Class	noun	verb	noun	verb
NOUN	55.7%	6.6%	86.9%	11.7%
VERB	19.7%	54.8%	49.4%	64.9%

statistically naive measures of similarity and category assignment, and the high levels of noise present in the sample, the proportion of successful classifications is surprisingly high. It is possible that this technique is capable of being refined in order to improve the accuracy achieved.

Conclusion

The results presented here demonstrate the feasibility of using simple, locally computed statistics of the speech to which children are exposed, in order to bootstrap syntactic categories. It is likely that, especially for languages where word order is not closely constrained, other sources of information (intonation, stress, language external factors *etc.*) are involved in this process. It appears plausible that a variant of the method could be applied at a lower level to languages such as Turkish, where word order is believed to play little role in early acquisition, being supplanted by highly regularised case markers. (Bates and MacWhinney, 1987). Statistical methods in general may play a large role in early language acquisition, via processes analogous to the one described here.

Future work. A quantative measure of the 'goodness' of clustering/categorisation is obviously required. This would ideally allow comparison of results independently, to some extent, of sample size, and hopefully across languages.

The output at each linguistic level (word, phrasal *etc.*) might be utilised in 'preprocessing' the input to the level above, ideally within a single, coherent process. Additionally, the growth and 'solidification' of categories, as data accumulates, and any correspondence with observed developmental stages is of interest.

Experiments with multi-dimensional scaling have suggested that the distance information implicit in the 600-dimensional space of possible distributions of context can in fact be adequately represented in many fewer dimensions. It may be possible to exploit this redundancy in amassing context information, and/or it may prove an interesting measure of the complexity of natural language. Similarly, the present method in itself, constitutes a generally revealing tool for the study of linguistic structure.

The most obvious remaining question is the performance of this method with other natural languages, particularly those whose word order is less constrained than English, *e.g.* Turkish, Japanese. As well as motivating the collection of the data that will be required, the thorough application of such methods to a variety of languages is likely to necessitate the integration of information from multiple sources, both distributional and otherwise, hopefully leading to more general and psychologically valid, models of the category acquisition process.

References

Bates, E. & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*, Hillsdale, NJ: LEA.

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

Chater, N. & Conkey., P. (1992). Finding linguistic structure with recurrent neural networks. *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, 420–407. Hillsdale, NJ: LEA.

Chomsky, N. (1980). *Rules and Representations.* Cambridge, MA: MIT Press.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, **14**:179–211.

Finch, S. & Chater, N. (1991). A hybrid approach to learning syntactic categories. *Artificial Intelligence and Simulated Behaviour Quarterly*, **78**:16–24.

Finch, S. & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society*, 820–825. Hillsdale, NJ: LEA.

Finch, S. & Chater, N. (1993). Learning syntactic categories: A statistical approach. In M.R. Oaksford and D.A. Brown, editors, *Neurodynamics and Psychology.* London: Academic Press.

Hanson, S.J. & Kegl, J. (1987). Parsnip: A connectionist model that learns natural language from exposure to natural language sentences. *Proceedings of the Ninth Annual Meeting Of the Cognitive Science Society.* Hillsdale, NJ: LEA.

Kneser, R. & Hey, H. (1991). Forming word classes by statistical clustering for statistical language modelling. *Proceedings of QUALICO 1*, Trier, Germany.

Kuczaj, S.A. (1982). On the nature of syntactic development. In S.A. Kuczaj (ed.) *Language development, Volume 1: Syntax and semantics.* Hillsdale, NJ: LEA.

Lasnik, H. (1989). On certain substitutes for negative data. In W. Demopolous and R. May, editors, *Learnability and Linguistic Theory.* Dordrecht, Netherlands: Reidel.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, **12**:271–295.

Maratsos, M. (1988). The acquisition of formal word classes. In Y. Levy, I.M. Schlesinger & M.D.S. Braine (eds.), *Categories and Processes in Language Acquisition*, Hillsdale, NJ: LEA.

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

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

Scholtes, J.C. (1991). Using extended feature maps in a language acquisition model. In *Proceedings of the 2nd Australian Conference on Neural Networks*, January, 1991.