Addressing the Learnability of Verb Subcategorizations with Bayesian Inference

Mike Dowman (Mike@cs.usyd.edu.au)
Basser Department of Computer Science, F09,
University of Sydney, NSW2006, Australia

Abstract

Elman (1993) has shown that simple syntactic systems can be learned solely on the basis of distributions of words in text presentation. However Pinker (1989) has proposed that children must make use of verbs' semantic representations in order to infer their syntactic subcategorizations (semantic bootstrapping). Results reported here demonstrate how Bayesian statistical inference can provide an alternative, and much simpler, account of how subcategorizations are learned. The acquisition mechanism described here suggests that syntactic acquisition may involve a much larger component of learning, and less innate knowledge, than is presumed within mainstream generative theory.

Introduction

This paper investigates how children learn their first language, and in particular the syntactic system of that language. It conceives of the problem in the following way: when exposed to utterances in that language, how is it possible to infer the grammatical system which produced those utterances. Further, the learner is assumed not to know the meanings of the words, have access to prosodic cues to structure, or to receive feedback about which sentences are not grammatical.

Currently the major paradigm within which language acquisition is explained is the parameter setting framework (Chomsky, 1995). Within this framework it is proposed that knowledge of language is largely specified innately, and learning consists of identifying word tokens and setting a limited number of parameters according to the syntactic structures to which the child is exposed. Chomsky argues that this position is necessary because 'even the most superficial look reveals the chasm that separates the knowledge of the language user from the data of experience.' (p. 5).

Gold (1967) investigated this problem more formally, and proved that without negative evidence (explicit information about which sentences are ungrammatical) languages are not 'learnable in the limit' unless the class of languages which the learner may consider is restricted *a priori*, for example by innate knowledge. Below I will discuss an alternative result by Feldman, Gips, Horning and Reder (1969) which suggests that Gold's result is not relevant to the circumstances under which children learn languages.

Redington, Chater and Finch (1998) investigated to what extent syntactic categories could be inferred based on distributions alone, without knowing *a priori* what syntactic categories existed in the language. They formed vectors by taking the two preceding and two following context words for each occurrence of each target word in a large corpus of transcribed speech, and recorded how often each context word occurred in each position. Only the 150 most frequent words were used as context, and so this resulted in 600 dimensional vectors for each word (there being one entry for each of the 150 context words in each of four positions). Clustering those words whose vectors were most similar in terms of Spearman's rank correlation resulted in clusters which corresponded to appropriate word classes for most of the 1,000 target words. While this system was good in that it could be applied to naturally occurring speech, it was necessary to decide at what level of dissimilarity to form separate classes, and so it doesn't completely solve the problem of recovering the syntactic classes used by the original speakers.

Elman (1993) demonstrated that not only word classes, but also syntactic patterns in which words belonging to those classes appeared, could be learned without much innate syntactic knowledge, at least for simple languages. He trained a recurrent neural network to predict the following word in artificially generated sentences conforming to a simple syntactic system containing 23 words, and syntactic features such as number agreement and recursion in relative clauses. Once trained on 50,000 sentences in this simple language, the network performed at near optimum accuracy at predicting the subsequent word at any stage in a sentence, showing that the network had internalized the structural constraints implicit in the data.

While both Redington et al (1998) and Elman (1993) demonstrate that much of syntactic structure can be learned by making statistical inferences based on the distributions of words, Pinker (1989) suggests that some aspects of syntax cannot be learned in this way. He proposes that, in order to determine verbs' subcategorizations in the absence of negative evidence, children must rely on complex innate rules combined with knowledge of the verbs' semantic representations.

Verbs such as *give* can appear in both the prepositional dative construction (1a), and the double object dative construction (1b), but there is a class of verbs such as *donate* which can only appear in the prepositional construction, (1c and 1d). However Gropen et al (1989) observe that, based on the alternation between (la) and (lb), children generalize this alternation to verbs such as *donate*, and so produce ungrammatical sentences such as (1d). They also demonstrated that when presented with novel, nonce, verbs in the prepositional construction, children will productively use them in the double object construction in appropriate contexts. However, ultimately children do learn which verbs cannot occur in the double object construction, and so we need a theory which can explain why children first make such gen-

eralizations, and then subsequently learn the correct subcategorizations.

(1) a. John gave a painting to the museum.
b. John gave the museum a painting.
c. John donated a painting to the museum.
d. *John donated the museum a painting.

While the main point of Pinker (1989) is that syntax cannot be learned from distributions alone, he acknowledges that the fact that certain syntactic structures do not occur could be used as indirect negative evidence that these structures were ungrammatical. However, he notes that children can neither consider that all sentences which they have not heard are not grammatical, and nor do they rule out all verb argument structure combinations which they haven't heard. He notes that it is necessary to identify 'under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical?' (p.14). The computational model presented in this paper is able to do just this, and so predict that a verb such as donate cannot occur in the double object construction, while at the same time predicting that a novel verb encountered only in the prepositional construction will follow the regular pattern and also appear in the double object construction.

Bayesian Grammatical Inference

Most work in syntactic theory assumes that grammars are not statistical, that is that they specify allowable structures, but do not contain information about how frequently particular words and constructions occur. However, if grammars were statistical, it appears that it would be much easier to account for how they were learned. Feldman et al (1969) proved that as long as grammars were statistical, and so utterances were produced with frequencies corresponding to the grammar, then languages are learnable. They note that proofs that language isn't learnable rely on the possibility of an unrepresentative distribution of examples being presented to the learner. While under Feldman et al's learning scheme it is not possible to be certain when a correct grammar has been learned, as more data is observed it becomes more and more likely that the correct grammar will be identified.

Feldman et al's proof uses Bayes' theorem, which relates the probability of a hypothesis given observed data to the *a priori* probability of the hypothesis and the probability of the data given the hypothesis. For a fixed set of data the best hypothesis is that for which the product of the *a priori* probability of the hypothesis and the probability of the data given the hypothesis is greatest. Feldman et al relate the probability of a grammar (seen as a hypothesis about language) to its complexity – more complex grammars are less probable *a priori*. As grammars are statistical, it is also possible to calculate the probability of the data given a grammar. This leads to an evaluation criterion for grammars where the complexity of a grammar is weighed off against how much data it has to account for, and how well it fits that data. A more complex grammar can be justified if it accounts for regularities in the data, but otherwise a simpler grammar will be preferred.

Minimum coding length provides an efficient implementation of Bayesian inference, using information theory (Shannon, 1948), which allows us to quantify the amount of information in a formal description of a grammar. The amount of information conveyed by an event (or symbol in a grammar) is equal to the negative logarithm of its probability. It is conventional to take logarithms to base two, resulting in the units of quantity of information being bits. Within this framework the best grammar is that which, together with a description of a corpus of data in terms of the grammar, can be specified using the least amount of information.

While Feldman et al (1969) showed that, given two or more grammars, it is possible to decide which is the best given a corpus of data, they did not show how these grammars could be created. For any reasonably complex grammar, the number of possible, but incorrect, grammars of equal or simpler complexity is so large that it is not plausible that a child could consider each in turn. However, in the next section, I describe computational models which are able to learn grammars by starting with a simple grammar, and then making small iterative changes which gradually lead towards the correct grammar. This avoids the need to consider every single possible grammar, and so allows grammars to be learned within a reasonable amount of time.

Computational Models of Syntactic Acquisition

Langley (1995) and Stolcke (1994) used simplicity metrics to learn simple syntactic systems, while Goldsmith (submitted) has applied this approach to the acquisition of morphology. Both Langley and Stolcke's systems produced similar results to those found by Dowman (1998) using the model described in the next section, although Langley's (1995) system did not incorporate considerations of how well the grammar fitted the data. It is shown below how Dowman's (1998) model was used to obtain new results concerning the acquisition of verb subcategorizations.

Description of Model

Dowman's (1998) model learned grammars for simple subsets of several languages, including the English data given in Table 1, which corresponds to the grammar given in Table 2. The only *a priori* knowledge of the structure of the corpus which was available to the model was implicit in the grammatical formalism with which grammars were specified. This formalism restricted the model to using binary branching or non-branching phrase structure rules, introducing each word with a non-branching rule, and using no more than eight non-terminal symbols. The non-terminal symbols were all equivalent arbitrary symbols, except that each grammar would contain one special symbol, *S*, with which each top down derivation would begin.

The frequency, and hence probability, with which each symbol (including words) appeared in the grammar was specified, and so the amount of information required to specify each symbol in a grammar could be calculated (using Shannon's (1948) information theory). A specification of a grammar would consist of a list of groups of three symbols, one for a rule's left hand side, and two for its right

hand side (a special null symbol being incorporated for use in non-branching rules). As the grammar was statistical, it was also necessary to record how often each rule was used in parsing the corpus. It was assumed that a fixed amount of information could be used to specify these probabilities, and so 5 bits of information was added to the evaluation of the grammar per rule. (The assumption of 5 bits of information is fairly arbitrary, but sufficient for the purposes described here.) The total cost of the grammar was the amount of information needed to specify each symbol in the grammar, and each rule's frequency.

Table 1: Data for English

John hit Mary	Ethel thinks John ran
Mary hit Ethel	John thinks Ethel ran
Ethel ran	Mary ran
John ran	Ethel hit Mary
Mary ran	Mary thinks John hit Ethel
Ethel hit John	John screamed
Noam hit John	Noam hopes John screamed
Ethel screamed	Mary hopes Ethel hit John
Mary kicked Ethel	Noam kicked Mary
John hopes Ethel thinks Mary hit Ethel	

Table 2: Grammar Describing English Data

$S \rightarrow NP\ VP$	$V_s \rightarrow$ thinks
$VP \rightarrow$ ran	$V_s \rightarrow$ hopes
$VP \rightarrow$ screamed	$NP \rightarrow$ John
$VP \rightarrow V_t\ NP$	$NP \rightarrow$ Ethel
$VP \rightarrow V_s\ S$	$NP \rightarrow$ Mary
$V_t \rightarrow$ hit	$NP \rightarrow$ Noam
$V_t \rightarrow$ kicked	

Given such grammars, the data was then parsed left to right, bottom up, with only the first parse found for each sentence being considered, and an ordered list of rules needed to derive the sentence obtained. This list allows us to make a probabilistic encoding of the data in terms of the grammar. Given the probabilities of the rules, and always knowing the current non-terminal symbol being expanded (starting with *S*, and always expanding the left most unexpanded non-terminal), it is only necessary to specify which of the possible expansions of that symbol to make at each stage. Hence, if a grammar accounts well for regularities in the data, little information will be required to specify the data. If a symbol can only be expanded by a single rule (such as *S* in the grammar above), then no information is necessary to specify that that rule is used.

By summing the amount of information needed to specify the grammar rules, the frequencies of those rules, and the data given that grammar, we obtain an evaluation for each grammar, with lower evaluations corresponding to better grammars. However, in order to complete the model of acquisition, it is necessary to describe the search mechanism that was used for generating and testing grammars.

The model started learning with a simple grammar of the form given in Table 3, with a rule introducing each word. This grammar is very simple, hence having a good evaluation itself, but it does not describe any regularities in the data, and so has a very bad evaluation in that respect, resulting in a poor overall evaluation.

Table 3: Form of Initial Grammars

$S \rightarrow X\ S$	$S \rightarrow X$
$X \rightarrow$ John	$X \rightarrow$ thinks
$X \rightarrow$ screamed	$X \rightarrow$ Ethel

The model would begin learning by making one of four random changes to the grammar, either adding a new rule (which would be the same as an old rule, but with one of the symbols changed at random), deleting a randomly chosen rule, changing one of the symbols in one of the rules, or the order of the rules, or adding a pair of rules in which one non-terminal symbol occurring on the left hand side of one and the right hand side of another was changed to a different non-terminal symbol. These changes are slightly simpler than those described in Dowman (1998), but further investigations have revealed that this learning system works well, and it was able to reproduce the results obtained with the more complex system, so it was used for deriving the new results presented in this paper.

After each change the evaluation of the new grammar with respect to the data would be calculated. If the change improved the evaluation of the grammar then it would be kept, but if the new grammar was unable to parse the data, it would be rejected. If the change made the evaluation of the grammar worse, then the probability that it would be kept would be inversely proportional to the amount by which it made the evaluation worse, and also throughout learning the probability that changes resulting in worse evaluations would be accepted was gradually reduced. This is an implementation of annealing search, which enables the system to learn despite finding locally optimal grammars in the search space. The program learned in two stages, in the first only taking account of the evaluation of the data in terms of the grammar (making it easier to find the grammatical constructions which best fitted the data), and in the second taking account of the overall evaluation (and so removing any parts of the grammar which could not be justified given the data). After a fixed number of changes had been considered (less than 18,000 in the case of the above data) learning would finish with the current grammar, no improvements usually having been found for a long time. For efficiency reasons, there were also limits placed on how deeply the parser could search for correct parses, and on the maximum number of rules which the grammar could contain at any stage of the search. Because the search strategy is stochastic, it is not guaranteed to always find the optimal grammar every time, so the learning mechanism would run the search several times, and select the grammar with the best overall evaluation.

Results

When used to learn from the English data in Table 1, the system learned a grammar which corresponded exactly to that in Table 2 in structure. (As linguistic categories are not known *a priori*, the system simply used a different arbitrary

symbol to represent each learned category.) Table 4 shows that this grammar was preferred because, while the grammar itself is more complex than the initial one, and so receives a worse evaluation, it captures regularities in the data, and so improves the evaluation of the data with respect to the grammar by a greater amount. Dowman (1998) used this same learning system (without any modifications except to the maximum number of non-terminal symbols) to learn aspects of French, Japanese, Finnish and Tigak.

Table 4: Evaluations for English Grammar

	Initial state of learning	Learned Grammar
Overall Evaluation	406.5 bits	329.5 bits
Grammar	160.3 bits	199.3 bits
Data	246.2 bits	130.3 bits

Learning Verb Subcategorizations

Given Dowman's (1998) success in learning simple syntactic systems, it was decided to investigate whether the same model could be used to learn some of the kinds of phenomena which it has been argued are especially problematic for theories of learning. In particular it was investigated whether the distinction between sub-classes of ditransitive verbs such as *gave* and *donated* could be learned.

There were three key results which the model aimed to replicate. Firstly, children eventually learn a distinction between verbs which can appear in both the double object and prepositional dative constructions, and those which do not show this alternation. Secondly, when children encounter a previously unseen verb they use it productively in both constructions. Finally, during learning, before children have seen many examples of an irregular verb which only occurs in a subset of the possible constructions of other verbs, they use that verb productively in constructions in which it is not grammatical.

Data Used for Learning

The same model was used as in Dowman (1998), but this time the data consisted of two types of sentences, prepositional datives such as (2a) and (2b), containing one of the verbs *gave, passed, lent,* or *donated*, and double object datives such as (2c), containing *gave, passed* or *lent*, but not *donated*. Each of these four verbs occurred with roughly equal frequency, and the alternating verbs were just as likely to appear in either construction. In addition the sentence (2d) was added, containing the only example of the verb *sent*. Noun phrases consisted of either one of two proper nouns, or one of the two determiners *a* or *the*, followed by either *painting* or *museum*. There were no biases as to which noun phrase was most likely to occur in which position, and overall the data consisted of 150 sentences.

No modifications were made to the model of Dowman (1998), except that in order to cope with the more complex data set the maximum number of non-terminals was increased to 14, and the number of iterations in the search was also increased.

(2) a. John gave a painting to Sam.
b. Sam donated John to the museum.
c. The museum lent Sam a painting.
d. The museum sent a painting to Sam.

Results

The initial and final evaluations of the grammars are given in Table 5. Again a more complex grammar has been learned which accounts better for regularities in the data than the original grammar. Examination of the learned grammar showed that the verbs had been divided into two classes (they have different symbols on the left hand sides of the rules producing them). *gave, passed, lent* and *sent* had all been placed in one class, while *donated* appeared in a class of its own. The grammar is able to generate only grammatical sentences, so *gave, passed, lent* and *sent* may appear in both double object and prepositional constructions, while *donated* may occur only in the prepositional dative construction. This has been learned even though there was no data explicitly indicating that *donated* did not follow the regular pattern, and even though *sent* only occurred once, and in the prepositional structure.

Table 5: Evaluations for Ditransitive Verbs Data

	Initial state of learning	Learned Grammar
Overall Evaluation	3445.6 bits	1703.4 bits
Grammar	190.3 bits	321.0 bits
Data	3255.3 bits	1382.3 bits

The results above account both for eventual learning of the distinction between syntactically distinct verbs such as *gave* and *donated*, and the productive use of novel verbs in regular constructions. The final phenomenon which we aimed to demonstrate was that, at earlier stages of learning, children overgeneralize and use verbs such as *donated* productively in constructions in which they are ungrammatical. In order to investigate this phenomenon, the total amount of data was reduced, to simulate a stage of acquisition where children had not been exposed to so many examples of each kind of verb. When the model learned from this data it failed to maintain a distinction between sub-classes of verbs, allowing all verbs to occur in both constructions. This was because there were not enough examples of *donated* to justify making the grammar more complex by creating a separate syntactic class, and so it was simply placed in the regular class.

Discussion

These results on the acquisition of regular and irregular verb subcategorizations show that an aspect of syntax is learnable which many other theories would have difficulty accounting for. In particular it is interesting to compare the performance of the model described here to that of connectionist models of syntactic acquisition such as Elman (1993).

Elman's network learned a language containing only 23 words, and yet 50,000 sentences were used to train the net-

work. This means that every word could have been observed in every syntactic position may times over, greatly reducing the need to form generalizations. Christiansen and Chater (1994) investigated to what extent this kind of model was able to generalize to predict that a word observed in one syntactic position would also be grammatical in another position. In order to do this, they trained a similar connectionist network on a more complex language containing 34 words, again using 50,000 sentences. In the training data they did not include *girl* and *girls*, in any genitive contexts, and, *boy* and *boys* in any noun phrase conjunctions. After training they found that the network was able to generalize so that it would allow *boy* and *boys* to appear in noun phrase conjunctions, but it didn't generalize to allow *girl* and *girls* to occur in genitive contexts. Christiansen and Chater considered the learning to have been successful in the case of *boy* and *boys*, but not in the case of *girl* and *girls*.

However, the account of the acquisition of verb subcategorizations presented in this paper relies on statistical properties of the data, and in particular the non-occurrence of certain forms. So, given 50,000 sentences of a language with only 34 words, in which two words did not appear in a given construction, it would seem that a learner would predict that this could not simply be due to chance. Given this perspective, it seems that Christiansen and Chater's network has learned correctly in the case of *girl* and *girls*, but not in the case of *boy* and *boys*.

In order to account for distinctions between *gave* and *donated*, it seems that neural networks must be more sensitive to quantitative information in language. The degree to which recurrent neural networks generalize is partly dependent on the fixed architecture of the network, and in particular on the number of hidden nodes. Bayesian learning methods for neural networks (MacKay, 1995) should be able to solve this problem, by placing a prior probability distribution on network structures and parameter values, although I am not aware of any applications of such networks to models of language acquisition.

Redington et al's (1998) system for learning word classes is capable of making very fine distinctions between subclasses of verbs, but unlike the system described here it is not able to decide when the distributions of two words are dissimilar enough that they should be placed into separate classes, and when the difference in distributions is simply due to chance variation within a class. However Boulton (1975) describes a program which does incorporate a Bayesian based metric into this kind of clustering system, and so demonstrates that it is possible to learn discrete classes automatically.

Certainly evaluation procedures based on simplicity metrics are not new to linguistic theory. Chomsky's (1965) theory of syntactic acquisition relied on such a measure to choose between alternative grammars. However, it is possible to identify some key differences which make Chomsky's theory very different to the Bayesian approach suggested here. Firstly Chomsky considered syntax to be fundamentally non-statistical. He had earlier argued that 'Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances….[P]robabilistic models give no particular insight into some of the basic problems of syntactic structure.' (Chomsky, 1957, p17). It seems hard to explain how any system which didn't monitor the frequencies with which verbs such as *donated* and *gave* are used would be able to account for how the different subcategorizations of these verbs could be acquired.

Probably an even more important difference between the kind of simplicity measure proposed in Chomsky (1965) and the kind used here, is that Chomsky did not incorporate a measure of goodness of fit to data into his simplicity metric. Chomsky's metric simply looked for the grammar which was shortest, in terms of the number of symbols which it contained. The theory relied on innate constraints on what forms grammar could take in order that 'significant considerations of complexity and generality are converted into considerations of length, so that real generalizations shorten the grammar and spurious ones do not.' (p42). Ultimately any notion of a simplicity metric was dropped from syntactic theory, because little progress seemed to be being made in understanding grammar selection in this way.

Interestingly however, Chomsky's (1965) theory shows that simplicity metrics are not necessarily incompatible with theories which postulate very strong innate constraints on grammar. It seems that even within a parameter setting model of language acquisition, statistical inferences would make the task of learning much easier, especially given the presence of noise in the data from which people learn (due primarily to grammatical errors, and exposure to data from children who have not mastered certain aspects of grammar). Showing that Bayesian inference can be useful in explaining language acquisition does not necessarily mean that it is actually used. Essentially it allows us to return the degree to which language is determined by innate principles of grammar to an empirical question, allowing the possibility of a much greater degree of learning in the process of syntactic acquisition.

However, postulating that a Bayesian mechanism is used in acquiring syntax results in very different predictions about what form syntactic knowledge will take than if we presume that language is largely determined by universal principles. Chomsky (1995) has argued that the language faculty of the mind should satisfy 'general conditions of conceptual naturalness that have some independent plausibility, namely, simplicity, economy, symmetry, nonredundancy, and the like' (p. 1). While Chomsky notes this is 'a surprising property of a biological system' (p. 5) he argues that this view is justified because throughout the history of syntactic research systems conforming to this kind of principle have turned out to be the right ones. However, if language is learned with a Bayesian system we would not expect it to conform to such principles. Grammars could contain a lot of irregular rules if these accounted well for regularities in observed language. Even the principle of lexical minimization is not so clear cut within a Bayesian based account of learning, as Bayesian metrics will favor grammars which associate a lot of information with individual words if this allows them to account better for regularities in the data. Hence, one prediction of Bayesian theory is that

the most commonly occurring words may be very idiosyncratic and irregular in their behavior, while very rare ones must conform to regular patterns.

It is interesting to compare the Bayesian account of acquisition of subcategorizations presented here to Pinker's (1989) theory. Pinker's theory predicts that universal innate principles relate the meaning of a word to its syntactic subcategorization. Instead of the syntactic subcategorization of a verb being determined empirically by a learner based on observations of patterns of occurrence, it is determined by the meaning of that verb. Certainly Gropen et al (1989) have shown that children are sensitive to correlations between semantic and phonological characteristics of verbs, and which subcategorization frames they are most likely to occur in. However, it is quite possible that these patterns were learned by the child in much the same way as we have proposed that syntactic subcategorizations may be learned. It would be interesting to investigate empirically whether children or adults could be influenced to prefer verbs in one construction or another by controlling the exemplars of these verbs to which they were exposed, perhaps by using artificial language experiments or nonce verbs integrated into natural languages. This kind of experiment should be able to resolve to what extent children make use of innate principles versus learning in determining verbs' subcategorizations.

The main limitation of the computational model described here is that it can only learn from small artificial data sets. There is no reason in principle why it cannot operate on naturally occurring language, it is simply that it would take an extremely long time to run on this kind of corpus. This is clearly a limitation which is shared with connectionist approaches, though Redington et al (1998) demonstrate impressive results learning from real language corpora. Current research is investigating ways in which the search procedure could be made more efficient, so that learning from more realistic corpora is possible, though it seems worth acknowledging that we are modeling a process which takes place over many years, and that the human brain is much more powerful than any computer.

Conclusion

This paper has shown that Bayesian inference is able to provide a simple and plausible account of how a number of aspects of syntax could be learned. In particular the computational model described here can learn verb subcategorizations where one verb is grammatical in only a subset of the structures in which another can appear, and yet predicts that newly encountered verbs are used productively in regular patterns. The model also accounts for overgeneralization and hence the use of irregular items in regular constructions during early stages of acquisition. While it is not logically necessary that children must make use of Bayesian inference in learning language, it has the potential to be incorporated into theories as diverse as recurrent neural networks and universal grammar.

Acknowledgments

I would like to thank Jeff Elman, David Powers, Brett Baker, Hong Liang Qiao, Adam Blaxter Paliwala, Cassily Charles, and two anonymous reviewers, for helpful comments on this paper. This research was supported by ARC and IPRS scholarships.

References

Boulton, D. M. (1975). *The Information Measure for Intrinsic Classification.* Ph.D. Thesis, Monash University, Melbourne.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The Minimalist Program.* Cambridge, MA: MIT Press.

Christiansen, M. H. & Chater, N. (1994). Generalization and Connectionist Language Learning. *Mind and Language, 9,* 273-287.

Dowman, M. (1998). *A Cross-linguistic Computational Investigation of the Learnability of Syntactic, Morpho-syntactic, and Phonological Structure* (Research Paper EUCCS-RP-1998-6). Edinburgh, UK: Edinburgh University, Centre for Cognitive Science.

Elman, J. L. (1993). Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition, 48,* 71-99.

Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical Complexity and Inference* (Tech. Rep. CS 125). Stanford, CA: Stanford University: Computer Science Department.

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

Goldsmith, J. (submitted). Unsupervised Learning of the Morphology of a Natural Language.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R. & Wilson, R. (1989). The Learnability and Acquisition of the Dative Alternation in English. *Language, 65,* 203-257.

Langley, P. (1995), *Simplicity and Representation Change in Grammar Induction* (Unpublished Manuscript). Palo Alto, CA: Institute for the Study of Learning and Expertise.

MacKay, D. J. C. (1995). Bayesian Methods for Supervised Neural Networks. In Arbib, M. A. (Ed.) *The Handbook of Brain Theory and Neural Networks.* Cambridge, MA: MIT Press.

Pinker, S. (1989), *Learnability and Cognition The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science, 22,* 425-469.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27,* 379-423 & 623-656.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models.* Doctoral dissertation, Department of Electrical Engineering and Computer Science, University of California at Berkeley.