Gobet, F., & Pine, J. M. (1997). Modelling the acquisition of syntactic categories. *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, p. 265-270. Hillsdale, NJ: Erlbaum.

Modelling the acquisition of syntactic categories

Fernand Gobet **Julian Pine**
ESRC Centre for Research in Development
Instruction and Training
Department of Psychology
University of Nottingham
Nottingham NG7 2RD
United Kingdom
{frg, jp}@psyc.nott.ac.uk

Abstract

This research represents an attempt to model the child's acquisition of syntactic categories. A computational model, based on the EPAM theory of perception and learning, is developed. The basic assumptions are that (1) syntactic categories are actively constructed by the child using distributional learning abilities; and (2) cognitive constraints in learning rate and memory capacity limit these learning abilities. We present simulations of the syntax acquisition of a single subject, where the model learns to build up multi-word utterances by scanning a sample of the speech addressed to the subject by his mother.

Introduction

This research represents an attempt to model the child's acquisition of syntactic categories. A computational model, based on the EPAM theory of perception and learning, is developed. The basic assumptions are that (1) syntactic categories are actively constructed by the child using distributional learning abilities; and (2) cognitive constraints in learning rate and memory capacity limit these learning abilities. The aim of the project is to build a distributional learning mechanism that is not only capable of constructing grammatical categories, but also of doing so in a way that is consistent with recent findings in the developmental literature on the sequencing of grammatical category acquisition.

Shortcomings of Cognitive-Semantic Constructivist Models

There has been a growing awareness in recent years of the shortcomings of constructivist models of grammatical development based on the gradual extension of broad cognitive-semantic categories. First, there is the problem that children's early grammatical knowledge does not appear to be restricted in the way that such models would seem to predict (e.g. Maratsos & Chalkley, 1980). For example, children tend to apply morphological markers (e.g. -ed past tense endings) to verbs whether those verbs refer to actions or not, while failing to overgeneralise the same markers to actional adjectives (e.g. 'Noisy', 'Naughty', etc.). Second, it can be shown that the kinds of broad cognitive categories to which such models typically appeal are often only viable as the semantic core of categories in a subset of the world's languages. For example, use of the semantic-cognitive category 'Agent' as a way of bootstrapping up to the category of NP subject would represent a false step in the acquisition of some ergative languages which carve up the semantics of Agency in a different way from nominative-accusative languages (e.g. Braine, 1988). Third, there is now a wealth of evidence that children are capable of acquiring linguistic distinctions which have little or no semantic base from very early in development. This includes evidence regarding the mass-count distinction in English (Gathercole, 1985), noun/verb distinctions in Hebrew (Levy, 1988), and linguistic gender in a variety of different languages (e.g. Karmiloff-Smith, 1979); and suggests that children may be sensitive to distributional properties of the language they are learning from a very early age.

Shortcomings of Nativist Accounts

The demise of semantic models of grammatical development has coincided with a resurgence of nativist accounts of children's early multi-word speech (e.g. Pinker, 1984; Valian, 1986; 1991). Such accounts typically use the semantic heterogeneity of children's early multi-word speech to argue for a more abstract level of analysis involving adult-like syntactic categories. However, although coherent in their own terms, these accounts do not fit the developmental data all that well and tend to have particular difficulty accounting for the lexical-specificity of children's early multi-word speech. Thus, as Braine (1988) points out, Pinker's attribution of a S -> NPsub + VP rule to young language learning children ignores the fact that children initially tend to order different NPsub + VP sequences in different ways. Similarly, Valian's attribution of a syntactic determiner category to young language-learning children hides lexical specificity in the contexts in which different determiners are used (Pine & Martindale, 1996; Pine & Lieven, in press); and her attribution of knowledge about nominative case-marking ignores the fact that there is typically no evidence for the contrastive use of case-marked pronouns in children's early multi-word speech (Lieven, Pine & Baldwin, 1997).

Perhaps the strongest challenge to nativist accounts of early multi-word speech, however, comes from Tomasello's work on the development of the verb category. Tomasello argues on the basis of evidence both from naturalistic multi-

word speech data (Tomasello, 1992) and from experimental studies (Olguin & Tomasello, 1993; Tomasello & Olguin, 1993), that there is a developmental asynchrony in the acquisition of the noun and verb categories in English. Thus, whereas even very young children show great facility in slotting novel noun-arguments into familiar verb structures, their knowledge about SVO word order seems to be lexically-specific in that they not only fail to generalise it from one verb to another, but also seem unable to use it as a cue for sentence comprehension, at least in the absence of additional supporting cues such as animacy and/or pronoun case-marking.

These findings are important for a number of reasons. First, they cast doubt on the validity of strong nativist accounts of children's early multi-word speech; second, they are consistent with a more gradual category-formation process which capitalises on the kind of distributional learning abilities required by the child to acquire semantically arbitrary categories such as linguistic gender; and third, they provide information about the developmental sequencing of early category acquisition which can guide the modeller by serving as a target for simulation and hence as a means of constraining the development of a viable distributional learning mechanism.

Distributional Learning and Word Class Acquisition

Distributional approaches to language have a long history in both Psychology and Linguistics. Moreover, recent work in this area has been quite successful in demonstrating just how much information is present in the statistical distribution of words in large text-based and conversation-based corpora (e.g. Finch & Chater, 1992). However, constructivist models based on such approaches have tended to be given rather short shrift in the language acquisition literature for a number of reasons. First, they have been criticised on logical grounds. Thus, certain formal properties of human languages, such as the presence of long-distance dependencies, have been seen as ruling out such accounts a priori. However, as recent work by Elman (1993) has shown, it is possible for a relatively simple distributional learning mechanism to learn such dependencies, albeit in a toy language, provided that analysis is initially restricted, either by phasing the input or by constraining the size of the mechanism's processing window. This suggests that such logical arguments should be treated with a certain amount of scepticism since they derive much of their power from the way in which they conceptualise language acquisition as a single logical problem rather than as a complex developmental process.

Second, distributional learning accounts have been criticised for making unrealistic assumptions about the child's processing abilities. Thus, Maratsos and Chalkley's (1980) model assumes that the child tabulates all the grammatical properties of all words and constituents in the input, so that the distributional analyser can flexibly find the 'best' features around which to build particular linguistic categories. However, as Keil (1981) points out, there are far too many possible 'grammatical properties' in the input for such an approach to be viable. Indeed, Pinker (1984) shows that even using an impoverished criterion for 'grammatical property' a seven-word sentence has about 9,000,000 possible 'grammatical properties'.

This is, of course, a very powerful argument against viewing language acquisition as a process of unconstrained distributional analysis. However, it is not an argument against a distributional approach to language acquisition *per se*. Moreover, it is worth pointing out that arguments against crediting the child with overly powerful distributional learning mechanisms are not the exclusive property of nativists and have actually been made on both sides of the nativist-empiricist divide. Thus Braine (1987) argues that one way of making progress in the area is to identify the limits of human distributional learning abilities and to use this knowledge as a constraint on the mechanisms proposed for natural language acquisition. According to Braine, experimental studies of artificial language-learning show that, under serial presentation conditions, subjects readily learn the positions of words or phrases with respect to a marker. However, they have great difficulty in learning arbitrary dependencies between classes of words. This, together with the results of Elman's work, raises the question of the extent to which it might be possible to use such information processing constraints to avoid some of the pitfalls faced by more powerful distributional learning mechanisms.

Modelling the Developmental Sequencing of Word Class Acquisition

Our modelling approach has several aims. First, we want to explore the effects of imposing constraints on a relatively simple distributional learning mechanism. These include constraints on learning rate and memory capacity (cf. Elman's work). Second, we want to build a learning mechanism which can closely simulate at least some aspects of the developmental data on word class acquisition, including the developmental asynchrony in the acquisition of noun and verb categories in English, and the verb-specific nature of children's early knowledge about SVO word order. Finally, we want to use this model to generate predictions about the precise shape of children's grammatical knowledge at particular points in development. The intention here is to use the predictions generated by the model to guide the analysis of a very detailed corpus of early multi-word speech data being collected in a parallel research project (Lieven & Pine, 1995).

As our computational framework, we have chosen EPAM, a theory that has a long history of successful simulation of human cognition, including: verbal learning behaviour (Feigenbaum & Simon, 1984), learning, memory and perception in chess (Simon & Gilmartin, 1973; Gobet 1993a,b; De Groot & Gobet, 1996), the digit-span task (Richman et al., 1995), and the context effect in letter perception (Richman & Simon, 1989). (For alternative approaches to modelling early language acquisition, see Langley, 1987, Reeker, 1976, or Selfridge, 1981).

EPAM models learning as the construction of a discrimination net. The basic mechanisms are as follows. During perception, an object is sorted through a sequence of tests, each related to some feature of the object. When the

description of the object mismatches the internal representation (the *image*) it has been sorted to, a new test is added in the tree, a test that relates to the mismatched feature. This mechanism is called *discrimination*. When the object is sorted to an internal representation that under-represents it, new features are added to the image by chunking. This mechanism is called *familiarisation*. EPAM specifies learning parameters based on empirical data from verbal learning experiments, such as 8 s to create a new node, 10 ms to carry out a test, as well as parameters on the size of short-term memory (STM), which is about 7 slots—an important question we are interested in in this research is the effect of varying this parameter on learning.

Until now, EPAM has been explored mainly as a theory modelling *access* to long-term memory (LTM). Simon (1989) has proposed that EPAM nets constitute an index to procedural and declarative memories, but has not given any details about how this should be implemented in a working computational model. One of the goals of our research is to show how procedural and declarative memories can be created by linking the nodes of the discrimination net. In order to give the system the capability of learning new concepts (in our case, grammatical categories), we will use both the notion of *template* (Gobet & Simon, 1996), which has been used to simulate experts' behaviour, and which is related to Tomasello's notion of verb-island, and the idea of *similarity links* created from the nodes of the discrimination net (Gobet, 1996).

Description of the Model

Components of the Model

The model consists of a set of nodes accessible through a discrimination net (see Figure 1), which can be joined together by similarity links (see below for a description of the way in which these links are created). When nodes are accessed by recognition, they are activated. The duration of activation depends on the total number of nodes currently active, which is set by the parameter *maximum-number-of-activated-symbols* (when few nodes are activated, they may correspond to short-term memory slots). Another parameter, *minimum-number-of-shared-links*, is used to test whether a similarity link should be created between two symbols.

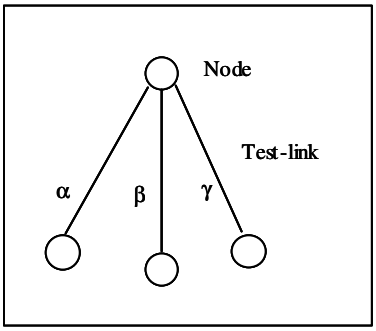

Figure 1: Illustration of the concepts of nodes and test-links in the discrimination net.

Learning Phase

During the learning phase, a corpus of utterances is presented in sequence to the program. The program learns the words and groups of words by discrimination and familiarisation, noting only relations of proximity and order, using the primitive "next." As learning proceeds, nodes grow tests for possible lexical items following them. When words (or groups of words) are recognised, the corresponding symbols are activated. Note that memory capacity, and therefore attention span, is counted in chunks, and not in syntactic units. Therefore, as the program learns, it will be able to augment the information contents of its activated memory as more chunks are recognised. We leave for future research the question of whether this knowledge-based increase will go in parallel with a "hardware" increase, with the number of slots in activated memory increasing as a function of time (see Ellis 1995, for similar ideas).

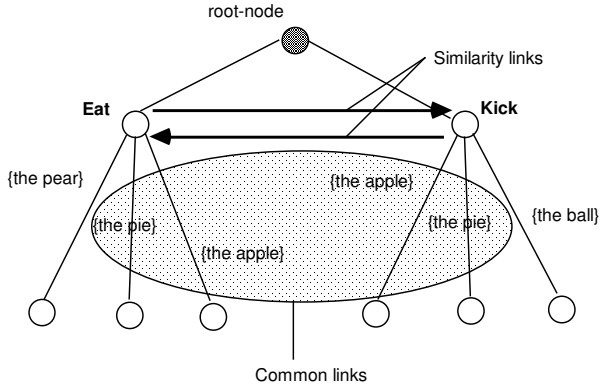

Figure 2: Creation of similarity links.

After a node is activated, the program tests whether there is another activated node that shares a number of similar test-links equal to or greater than the parameter *minimum-number-of-shared-links*. If this is the case, two similarity links are created between the two nodes, one starting from the first node to the second, the other starting from the second node to the first (see Figure 2).

Thus, frequency and variety of occurrences play a key role in the basic mechanisms we have outlined. We propose that lexically-specific 'Subject-Verb' chunks or 'Verb-Object' chunks will be formed in this way. Let's take the Verb-Object case as an example. In the first step,[1] the individual words are learned. In the second step, nodes grow tests for possible lexical items following them. Later (third step), when two verb-nodes are activated, a comparison is made of the attributes of each of these verb-nodes. If it is found that the number of test-links common to both nodes is larger than the *minimum-number-of-shared-links* parameter, a similarity link is created between the two nodes. Testing for similarity in the test-links ensures that

[1] Note that the individual steps overlap.

similarity links are generally created only between nodes representing words of the same syntactic class, even though the program itself does not know these classes. Finally, the third process is repeated and leads to the creation of networks of links. These networks, which may be knowledge islands isolated from each other, can later be used by the program to generate 'verb+object' sequences (see "Performance phase"). Note that memory capacity limits the probability of learning such links.

Performance Phase

During the performance phase, the model produces utterances in two ways: (a) by *recognition:* the model outputs the image of a node and one of the test-links (this may be the symbol NIL, indicating that nothing follows the image); or (b) by *generation:* the model outputs the image of a node and one of the test-links of a node linked by a similarity link. Note that, if enough learning has taken place, the output of the image may be a rather complex utterance, for example: (GOING TO PUT HIM IN THE BOX)**.**

Simulations

Methods

We now present a simulation of the early syntactic development of a single subject, Richard, between the ages of 1;9 and 2;3 years. As a simulation of the parental input, we use a sample of his mother's speech, coded orthographically in CHILDES format,[2] containing 5630 utterances (some utterances are duplicated in the sample), recorded over a period of 16 months**.** For Richard's data, we use a sample of 789 bigrams, coded as types (i.e. all the bigrams occurring in Richard's corpus of 610 multi-word utterances).

The program learns by scanning the maternal sample. Each utterance is sorted through the discrimination net, and, when parts of the utterance are recognised, the corresponding symbol is activated. When the number of symbols activated is more than the *maximum-number-of-activated-symbols* parameter, the symbol activated for the longest time is deactivated. After a new symbol is activated, the program tests whether there is another activated symbol which shares at least *minimum-number-of-shared-links* similar links with it. When this is the case, a link is created between the two symbols. This link may be used later to generate sentences, as described earlier.

Results

In these simulations, we were interested in the role played by the *maximum-number-of-activated-symbols* and *minimum-number-of-shared-links* parameters. Since the performance of the program changes little after the first run through the sample (typically, the programs gains only 2-3% with the total number of sentences generated in the six next runs), we consider only the results after the first run here.

Acquisition of Nodes and Similarity Links. The acquisition of nodes by the discrimination and familiarisation mechanisms is not affected by the two parameters in question. In all cases, 7851 nodes were learned. In the test phase, the program was able to recognise all individual words from the mother's sample, but missed 10% of the words from the child's sample. The utterances containing any of these words were excluded from the following analyses.

Figure 3 shows that the number of similarity links is sensitive both to the maximum number of nodes activated and the minimum number of shared links. The effects of the two variables are additive. In the worst case (*maximum-number-of-activated-symbols* = 10 and *minimum-number-of-shared-links* = 3) only 68 links are created.

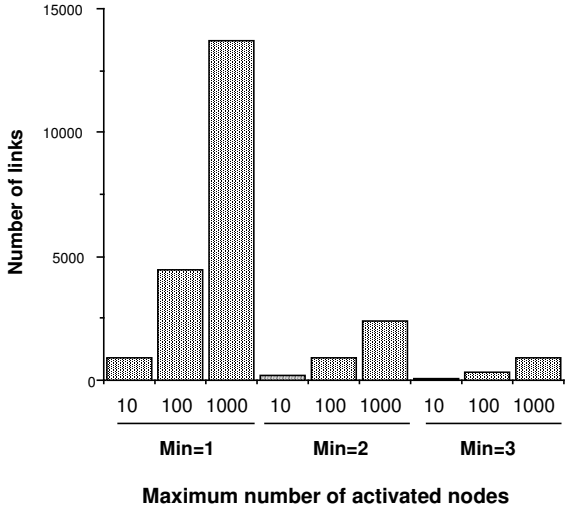

Figure 3: Number of similarity links that are learned by the program, as a function of the maximum number of nodes activated and the minimum number of shared test-links to create a similarity link.

Test with the Mother's Sample (Teaching Set). In all cases, 74% of the utterances were recognised as such by the program. Figure 4 illustrates the proportion of utterances generated by the program. As can be seen, both parameters are important. There is no indication of an interaction.

Test with the Child's Sample (Testing Set). In all cases, 34% of the utterances were recognised by the program. Figure 5 illustrates the proportion of utterances generated by the program. The results are similar to those observed with the mother's sample, with the qualification that more utterances were recognised and less generated with the latter sample.

[2] For simplicity sake, we will assume that phonological segmentation has already been done, as is common in many current theories of syntax acquisition. In principle, phonological segmentation can be obtained with the set of mechanisms included in the model described here.

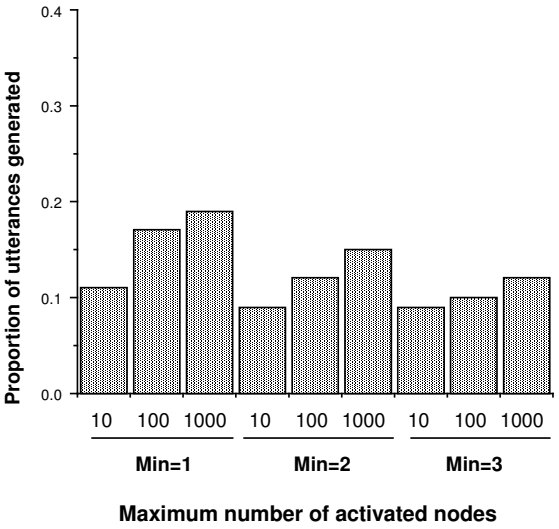

Figure 4: Proportion of the utterances from the mother's sample that are generated by the program, as a function of the maximum number of nodes activated and the minimum number of shared test-links to create a similarity link.

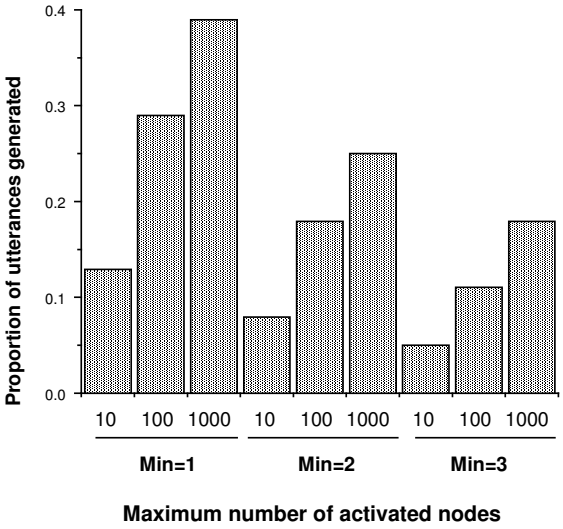

Figure 5: Proportion of the utterances from the child's sample that are generated by the program, as a function of the maximum number of nodes activated and the minimum number of shared "links" to create a similarity link.

Qualitative Analysis. A qualitative analysis of those of Richard's bigrams generated by the program and those which the program fails to generate illustrates the following properties of the model. First, it is highly sensitive to lexical patterns such as 'a + X' (20 instances), 'where + X' (16 instances) and 'X + gone' (13 instances); second, it is able to generate instances of these patterns to which it has not been previously exposed; and third, its performance on these patterns is not an all-or-nothing affair. Thus, although the model's generation of 'a + X' sequences suggests that it has picked up something like a 'indefinite article + Noun' pattern from the input, it actually fails to generate some of the 'indefinite article + Noun' bigrams generated by Richard himself. While in one sense this failure is obviously a weakness of the model, it does underline the fact that the model is building syntactic knowledge in a piecemeal fashion which is highly sensitive to the similarity structure of the input. This kind of piecemeal learning is consistent with a number of recent studies that suggest that although children are sensitive to the distributional properties of the input, their distributional knowledge is much more limited in scope that would be assumed by a traditional syntactic analysis (e.g., Braine, 1988; Tomasello, 1992; Pine & Martindale, 1996). It also suggests that the model has the potential to serve as a means of developing and testing hypotheses about the nature of relation between the structure of the input and the shape of the categories being learned.

What are the qualitative effects of the constraints imposed by the two parameters we have manipulated in our simulations? As shown earlier, using weak constraints (i.e., large number of activated nodes and small minimum number of shared links) allows the program to reproduce up to 74% of the child's bigrams (34% by recognition, and 40% by generation). The cost is that it generates a large quantity of non-grammatical utterances. For examples, when given as input "I" and asked to generate sentences, the model outputs things like (I WRONG) or (I ROAD), among some correct 'I + verb' sentences. By contrast, when the constraints are high (*maximum-number-of-activated-symbols* = 10 and *minimum-number-of-shared-links* = 3), the program generates only correct 'I + verb' sentences. As with Elman's simulations, this suggests some benefits of having a highly constrained cognitive system.

Conclusion

The computational approach to modelling syntax acquisition outlined above has several advantages over other work in this area: 1) Learning does not require feedback; 2) Nodes of the discrimination net and the semantic links between them are not defined in advance, but are dynamically created as a function of the input and of the current state of the system; 3) Since large discrimination nets and semantic memories can be developed, the model is not limited to "toy" domains; 4) All learning mechanisms are local, and respect psychological constraints; 5) The model allows us to study the role of memory capacity directly, a question that has mainly been addressed experimentally until now. In addition, the simulations described above suggest that constraints on memory capacity and minimum number of shared links may be necessary to ensure that the program is able both to generate and to restrict itself to grammatical sentences.

Our intention is to develop this model further in the future and use it not only to simulate multi-word speech data which have already been collected, but also to generate further hypotheses about the shape of children's grammatical knowledge at particular points in development. These will then be used to guide the analysis of a very detailed

multiword speech corpus currently being collected in a large-scale project in Nottingham and Manchester.

References

Braine, M. D. S. (1987). What is learned in acquiring word classes: a step towards an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Braine, M. D. S. (1988). Review of 'Language learnability and language development' by S. Pinker. *Journal of Child Language*, 15, 189-199.

de Groot, A. D. & Gobet, F. (1996). *Perception and memory in chess: Studies in the heuristics of the professional eye.* Assen: Van Gorcum.

Ellis, N. C. (1995). Sequence learning and language acquisition. *Studies in Second Language Acquisition*, 18, 91-126.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, 48, 71-99.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, *8*, 305-336.

Finch, S. & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Gathercole, V. (1985). 'He has too much hard questions': the acquisition of the linguistic mass-count distinction in much and many. *Journal of Child Language*, 12, 395-415.

Gobet, F. (1993a). *Les mémoires d'un joueur d'échecs*. [The memories of a chessplayer]. Fribourg: Editions Universitaires.

Gobet, F. (1993b). A computer model of chess memory. *Proceedings of 15th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Gobet, F. (1996). Discrimination nets, production systems and semantic networks: Elements of a unified framework. *Proceedings of the 2nd International Conference on the Learning Sciences*. Evanston, Ill.: Northwestern University.

Gobet, F. & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology*, 31, 1-40.

Karmiloff-Smith, A. (1979). *A functional approach to child language: a study of determiners and reference*. Cambridge: Cambridge University Press.

Keil, F. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, 88, 197-227.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley & R. Neches (Eds.), *Production system models of learning and development*. Cambridge, MA: The MIT Press.

Levy, Y. (1988). The nature of early language: Evidence from the development of Hebrew morphology. In Y. Levy, I. M. Schlesinger & M. D. S. Braine (Eds.), *Categories and processes in language acquisition*. Hillsdale, NJ: Erlbaum.

Lieven, E. V. M., & Pine, J. M. (1995). The developmental of grammatical categories. ESRC Project No. R000236393.

Lieven, E. V. M., Pine, J. M. & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24,* 187-219.

Maratsos, M. P. & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's Language.* (Volume II). New York: Gardner Press.

Olguin, R. & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.

Pine, J. M. & Lieven, E. V. M. (in press). Slot-and-frame patterns and the development of the determiner category. *Applied Psycholinguistics*.

Pine, J. M. & Martindale, H. (1996). Syntactic categories in the speech of young children: the case of the determiner. *Journal of Child Language*, 23, 369-395.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Reeker, L. H. (1976). The computational study of language acquisition. In M. Yovits & M. Rubinoff (Eds.), *Advances in computers.* New York: Academic Press.

Richman, H. B., & Simon, H. A. (1989). Context effects in letter perception: Comparison of two theories. *Psychological Review*, 96, 417-432.

Richman, H. B., Staszewski, J., & Simon, H. A. (1995). Simulation of expert memory with EPAM IV. *Psychological Review,* 102, 305-330.

Selfridge, M. (1981). A computer model of child language acquisition. *Proceedings of the Fourth International Joint Conference on Artificial Intelligence* (pp. 281-287).

Simon, H. A. (1989). *Models of thought.* (Volume II). New Haven: Yale University Press.

Simon, H. A., & Gilmartin, K. J. (1973). A simulation of memory for chess positions. *Cognitive Psychology*, 5, 29-46.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge: Cambridge University Press.

Tomasello, M. & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8, 451-464

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, 22, 562-579.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40, 21-81.