Shultz, T. R., & Bale, A. C. (2001). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy*, *2*, 501-536.

This version has some minor corrections concerning the proportion of reversals in the infant data. It was not made from page proofs or from the journal article, so page numbers and some other details may not correspond to the version published in *Infancy*.

Neural Network Simulation of Infant Familiarization to Artificial Sentences: Rule-like Behavior Without Explicit Rules and Variables

Thomas R. Shultz

Department of Psychology
McGill University

Alan C. Bale

Department of Linguistics
McGill University

Abstract

A fundamental issue in cognitive science is whether human cognitive processing is better explained by symbolic rules or by sub-symbolic neural networks. A recent study of infant familiarization to sentences in an artificial language seems to have produced data that can only be explained by symbolic rule learning and not by unstructured neural networks (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). Here we present successful unstructured neural network simulations of the infant data, showing that these data do not uniquely support a rule-based account. In contrast to other simulations of these data, these simulations cover more aspects of the data with fewer assumptions about prior knowledge and training, using a more realistic coding scheme based on sonority of phonemes. The networks show exponential decreases in attention to a repeated sentence pattern, more recovery to novel sentences inconsistent with the familiar pattern than to novel sentences consistent with the familiar pattern, occasional familiarity preferences, more recovery to consistent novel sentences than to familiarized sentences, and extrapolative generalization outside the range of the training patterns. A variety of predictions suggest the utility of the model in guiding future psychological work. The evidence, from these and other simulations, supports the view that unstructured neural networks can account for the existing infant data.

Supplementary materials to this article are available on the World Wide Web at http://www.infancyarchives.com

Requests for reprints should be sent to Thomas R. Shultz, Department of Psychology, McGill University, 1205 Penfield Avenue, Montreal, PQ, Canada H3A 1B1. E-mail: thomas.shultz@mcgill.ca

One of the most fundamental issues in contemporary cognitive science concerns the appropriate level of theoretical analysis of human cognition. Is human cognition a symbolic rule-based system or a sub-symbolic neural network system? Symbolic rules are if-then propositions that typically contain variables that can be bound to values (Anderson & Lebiere, 1998; Newell, 1990). For example, "If the goal is to use an English verb that describes the past, then add the suffix -ed to the stem of the verb." The key variable in this rule is *verb*, which can be bound to any of a very large number of English verbs. In contrast, artificial neural networks perform computations that are roughly inspired by brain processes. These computational characteristics include modulation of neuronal activity due to summation of excitatory and inhibitory impulses, modification of synaptic connections due to learning, and layers of neuronal connectivity (Hertz, Krogh, & Palmer, 1991). Such networks have been shown, for instance, to be capable of simulating not only the rule-like adding of suffixes to verb stems but also the exceptions found in irregular English verbs (Plunkett & Marchman, 1993, 1996).

A recent study of infant familiarization to sentences in an artificial language is portrayed as having struck a damaging blow to the neural network view by presenting data that can only be explained by rules and variables (Marcus et al., 1999). In that study, 7-month-old infants attended longer to sentences with unfamiliar syntactic structures than to sentences with familiar syntactic structures. A variety of experimental controls and some unsuccessful neural network models allowed the authors to conclude that ordinary, unstructured neural networks cannot simulate these results and that infants by default must possess a rule-learning capability that is not available to such neural networks. A companion article suggested that rule learning, because it was demonstrated in infants so young, may be an innately provided capacity of the human mind distinct from associative learning mechanisms like those in neural networks (Pinker, 1999).

The Marcus et al. (1999) claims were interesting because if they were true, they could trigger a retreat from otherwise promising neural approaches. A standard way to decide between alternative theories in cognitive science is to implement computational models to determine which model captures the data in the most precise, principled, and parsimonious fashion. In this article, we present neural network simulations of the key features of the Marcus et al. experiments, showing that their infant data do not uniquely support a rule-based account.

We begin the article with a brief review of the psychological evidence and we summarize a rule-based interpretation of that evidence. We then discuss current psychological and neural network interpretations of familiarization phenomena before reviewing existing models of the Marcus et al. (1999) data and presenting our new simulations. As well as covering the infant data on differential recovery of attention, the simulations are extended to study generalization abilities and to determine the nature of the knowledge learned by the networks. We develop predictions for new psychological research. Finally, we address the issue of whether the infant data require a symbolic rule-based explanation in view of the many successful unstructured neural network simulations.

PSYCHOLOGICAL EVIDENCE

Marcus et al. (1999) reported three experiments in which 7-month-old infants were familiarized to three-word sentences in an artificial language and then tested on novel sentences that were either consistent or inconsistent with the familiar pattern. The design of these experiments is shown in Table 1. In Experiment 1, infants were familiarized to sentences with either an ABA pattern (e.g., *ga ti ga* or *li na li*) or an ABB pattern (e.g., *ga ti ti* or *li na na*). In each case, there

were 16 of these sentences, created by combining four A-category words (*ga*, *li*, *ni*, and *ta*) with four B-category words (*ti*, *na*, *gi*, and *la*). After infants became familiar with a sentence pattern, they were presented with two novel sentences that were either consistent or inconsistent with the familiar pattern. For infants familiar with the ABA pattern, the inconsistent patterns were of the ABB form. For infants familiar with the ABB pattern, the inconsistent patterns were of the ABA form. The novel A-category syllables were *wo* and *de*; the novel B-category words were *fe* and *ko*.

Table 1

Design of Marcus et al.'s (1999) Experiments

Procedure	Experiments 1 and 2		Experiment 3	
	Condition 1	Condition 2	Condition 1	Condition 2
Familiarize	ABA	ABB	ABB	AAB
Consistent	ABA	ABB	ABB	AAB
Inconsistent	ABB	ABA	AAB	ABB

Note. A and B refer to two different categories of monosyllabic nonsense words. From “Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables.” By T. R. Shultz and A. C. Bale. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (p. 459), 2000. Mahwah, NJ: Erlbaum. Copyright 2000 by the Cognitive Science Society, Inc. Adapted by permission.

The dependent measure in the test phase of these experiments was looking time. If the infant looked at a flashing light to the left or right, a test sentence was played from a speaker situated near that light. Each test sentence was played repeatedly until the infant either looked away or until 15 sec elapsed. Infants attended more to inconsistent novel sentences than to consistent novel sentences, indicating that they were sensitive to differences between the two types of sentences.

Experiment 2 had exactly the same structure except that the words were chosen more carefully so that phoneme sequences were different in the familiarization and test patterns. Experiment 3 involved the same A and B categories of words as did Experiment 2 but employed contrastive syntactic patterns that each duplicated a consecutive word: AAB versus ABB. The idea of Experiment 3 was to rule out the possibility that infants may have used the presence or absence of consecutively duplicated words to distinguish syntactic types. For example, ABB sentences had consecutively duplicated words (the second B word is consecutively duplicated), but ABA sentences did not. Consequently in Experiment 3, both sentence types had consecutively duplicated words.

Infants in all three experiments attended more to inconsistent than to consistent novel sentences. The issue addressed in our article concerns the proper theoretical account of this syntactic processing. Is this processing based on rules and variables or on the mechanisms employed in unstructured neural networks, namely unit activations and connection weights?

A RULE-BASED INTERPRETATION

Marcus et al. (1999) argued that these simple grammars could not be learned by statistical methods common to standard neural networks. In particular, they considered sensitivity to transitional probabilities, discrepancies from stored sequences, and counting event frequencies. Transitional probabilities would not work because the transitional probabilities for novel words

would be 0. Noting discrepancies from stored sequences would not work because both consistent and inconsistent test sentences differ from the familiar sentences. Counting the numbers of consecutively duplicated words would work for Experiments 1 and 2 but not for Experiment 3, in which both sentence types had consecutively duplicated words. Marcus et al. also tried some neural network simulations using Simple Recurrent Networks (SRN) pioneered by Elman (1990). Details of these unsuccessful simulations were not provided, but Marcus et al. suggested the basic problem was that the network coding techniques employed did not permit generalization to novel words.

Instead, Marcus et al. (1999) claimed that a rule-based model could cover their data: "We propose that a system that could account for our results is one in which infants extract algebra-like rules that represent relationships between placeholders (variables) such as 'the first item X is the same as the third item Y' " (p. 79). This example rule was presumably designed by Marcus et al. to account for sentences with an ABA pattern.

Marcus et al. (1999) did not report an implementation of a rule-based model to account for the infants' gradual familiarization with sentence patterns or an analysis of how rule learning may be used to compute familiarity. In computer simulations, such explicit rules are usually processed with so-called production system programs that (a) match rule conditions against the facts in a working memory buffer, (b) select a rule whose conditions are satisfied, and (c) fire that rule, thus producing fresh conclusions or actions (Anderson & Lebiere, 1998; Newell, 1990). A production system model of the Marcus et al. data has not yet been reported. It is critical in such rule-based processing systems that, if variables are used, all variable bindings are preserved and accessible to further computation, a criterion that is not met by standard neural networks.

Marcus et al. (1999) suggested that their infant data may be accounted for by so-called structured neural networks that implement explicit rules and variables in a neural style:

> The problem is not with neural networks per se but with the kinds of neural networks that are currently popular. These networks eschew explicit representations of variables and relations between variables; in contrast, some less widely discussed neural networks with a very different architecture do incorporate such machinery and thus might form the basis for learning mechanisms that could account for our data (pp. 79-80).

Later we review structured network models of this type that capture the basics of the Marcus et al. data.

PSYCHOLOGICAL THEORY OF FAMILIARIZATION

Familiarization techniques like those used by Marcus et al. (1999) have often been termed *habituation*, with the term *dishabituation* applied to the recovery of attention to novel stimuli. A dominant psychological analysis of habituation assumes that infants gradually construct representational categories for stimuli that they encounter (Cohen, 1973; Sokolov, 1963). Infants tend to ignore stimuli that they have built categories for, while concentrating on stimuli that are relatively novel. This seems adaptive in encouraging infants to learn about the world. The technique of familiarizing infants with a type of stimulus during an experimental session and testing their reactions to stimulus changes has enabled dozens of discoveries of perceptual and cognitive abilities in young infants over the past 30 years (e.g., Cohen, 1979; Haith, 1990; Oakes & Cohen, 1990; Quinn & Eimas, 1996).

Such habituation and dishabituation phenomena are often discussed in terms of recognition memory. If there is substantial recovery of attention to a novel test stimulus, then it is

viewed as being novel. However, if there is little or no recovery of attention to a test stimulus, then it has been recognized as a member of a familiar category.

The course of familiarization is typically characterized by an exponential decrease in attention and processing. The decrease is gradual but starts at a fast rate that slows as the asymptote of no attention is approached. The decrease in attention is gradual because building representations takes time and effort for relatively naive infants. The slowing of the decrease is a natural consequence of the fact that attention typically starts at a high level and is bounded by zero attention at asymptote. Apparent deviations from this exponential decrease in attention may result from cutting off the familiarization phase before a stimulus category is fully formed.

NEURAL APPROACHES TO FAMILIARIZATION

Neural network techniques for simulating familiarization have been available for some time (Kohonen, 1988), but it is only recently that they have been applied to infant experiments. So-called encoder neural networks that learn to reproduce their inputs on their output units are capable of simulating familiarization and novelty effects in human infants (Mareschal & French, 1997). Relations among stimulus features are abstracted in hidden unit representations as connection weights are adjusted. The accuracy of these hidden unit representations is tested by expanding them onto the output units. Discrepancy between output and input representations is measured as network error. Stimuli that produce little or no error are effectively recognized as familiar. Stimuli that produce large error can be considered novel and deserving of further processing via weight adjustment. If there are fewer hidden units than input or output units, the encoder network learns to abstract a compact representation of the problem on its hidden units. Such abstractions enable construction of prototypes, generalization to novel inputs, and pattern completion (Hertz et al., 1991). Unlike earlier neural approaches to habituation (Kohonen, 1988), encoder networks can cope with nonlinear relations in the stimulus patterns and are not restricted to novelty detection. That is, they indicate not only whether a stimulus is novel but also its expected qualities.

A fundamental assumption in this modeling of familiarization with encoder networks is that interest reflects network error. A link between error and focus of learning is natural for neural networks for several reasons. First, neural networks learn through adjustment of their connection weights. Second, the sizes of weight changes are proportional to the first derivative (slope) of the function relating weight to error. Third, because these slopes are themselves proportional to error, weight change is proportional to error: the larger the error, the greater the weight change. Thus, networks have a natural tendency to focus learning on the largest sources of error in the training patterns. The psychological analog to this would be a tendency for greater focus on stimulus patterns that produce more error, namely those patterns that are least familiar.

PREVIOUS NEURAL NETWORK MODELS

There are currently eight alternative computational models of the Marcus et al. (1999) data, all of them connectionist models, presumably attracted by the challenge that ordinary connectionist models would be unable to simulate these data. Each of the eight models is able to distinguish or represent the difference between consistent and inconsistent sentences, which was the basic finding of the Marcus et al. experiments. Because complete information about some of these models is not yet available, we only briefly summarize each model here. All but two of the eight models are conventional, unstructured connectionist models without explicit rules and variables. Four of these unstructured models use the SRN architecture, construing the task of the network to be one of

predicting the next word or phoneme in the sentence when given a word or phoneme as input. SRNs are feed-forward networks with recurrent connections from the hidden units back to the input units. These recurrent weights allow a network to process sequential stimuli, such as sentences, by implementing a short-term memory for just-processed information.

Negishi (1999a, 1999b) used an SRN without hidden units. Each word is coded for both place of articulation of the consonant and vowel height. The strength of this model is that it has no hand-wired assumptions nor does it need any preexperimental experience to cover the Marcus et al. (1999) results. Marcus (1999a) suggested that this network implements a form of variable binding by using continuous values on the input units that are transmitted directly to the output units. However, this is a highly idiosyncratic interpretation of variable binding. There is no direct implementation of symbolic rules in Negishi's (1999b) network; there are only learnable connection weights and transitory unit activations. The computational style is clearly that of ordinary, unstructured neural networks.

Elman (1999) used an SRN with hidden units that encoded each word using binary phonetic features. The network was pretrained to distinguish whether each word differed from a previous word. This pretraining was motivated by a criticism that the SRNs used by Marcus et al. (1999) failed to simulate the infant data because they lacked normal phonemic experience (Seidenberg & Elman, 1999). However, it is unclear whether the type of pretraining used by Elman (1999) is psychologically realistic. Although 7-month-olds have experience with phonemes, it is not established that they receive explicit information about phonemic sameness and difference. Also, the task of this network was one of discrimination rather than the infants' task of habituation.

Christiansen and Curtin (1999) also used a pretrained SRN, but in this case the network was pretrained on word segmentation. Using binary phonological features for the inputs and outputs, the task of the network was to predict the identity and stress value of the next phoneme given a phoneme marked for stress and word boundary information as input. Then the network was trained on the Marcus et al. (1999) habituation sentences. When presented with the test sentences, the network was better at predicting phonemes occurring in inconsistent sentences than in consistent sentences. Although the network distinguished inconsistent from consistent test sentences, it is unclear why it would perform better on inconsistent ones.

The last of the four SRN models is that of Altmann and Dienes (1999). Their model did not require any pretraining and did simulate a habituation task rather than a discrimination task. However, Marcus (1999b) criticized this model for using somewhat unconventional correlation and distance measures to demonstrate that the network distinguishes between consistent and inconsistent sentences. If the prediction of the network is measured by the most active unit, as is relatively common, Marcus (1999b) claims that the network does not actually learn the training sentences but rather oscillates between the two grammars.

Gasser and Colunga (1999) simulated the Marcus et al. (1999) data using a specially-designed network with microrelation units. Connections were handwired in this network to manipulate input activations. As a result, similar syllables are synchronized in the network, which causes low activations on the microrelation units; dissimilar syllables are desynchronized in the network, which causes high activations on the microrelation units. This network requires no pretraining to achieve success on the habituation task, but the hand-wired use of temporal synchrony may be construed as implementing a form of explicit variable binding.

Shastri and Chang (1999; Shastri, 1999) designed a structured connectionist model that implemented explicit variable binding. This network had units representing sequential positions of words in three-word sentences and it coded the words with arbitrary binary features. Using explicit external feedback on word positions for three-word sentences, the network learned, for example, to represent the ABA pattern by firing the first position unit synchronously with the third position unit. Although this network generalizes well to any novel three-word sentences, it is extensively handbuilt and requires unrealistic feedback signals about the positions of words in a sentence.

Sirois, Buckingham, and Shultz (2000) applied a simple auto-associator network model to the Marcus et al. (1999) data. An auto-associator consists of a single layer of interconnected units, allowing internal circulation of unit activations over multiple time cycles. After learning the habituation sentences, these networks required more processing cycles to learn inconsistent than consistent test sentences. This model requires no handwiring of weights and no pretraining, and the mapping of processing cycles to recovery from habituation seems particularly natural. Furthermore, occasional reversals of preference were found as in the infant data. Because the auto-associator does not use hidden units, it would appear to be limited to learning only linearly separable patterns.

The final simulation is that of Shultz (1999), which modeled the habituation data with an encoder version of the cascade-correlation algorithm. There was an arbitrary analog coding of syllables. As in the model of Sirois et al. (2000), the task of the network was construed as one of recognizing whole three-word sentences. After habituating to the training sentences, the networks produced less error when processing consistent sentences than inconsistent sentences. As well as successfully simulating the Marcus et al. (1999) data, the networks demonstrated an exponential decrease in error during training, as is customary in infant habituation, and showed occasional reversals of preference as found with the infants in the Marcus et al. experiments. However, the use of arbitrary analog coding is not psychologically realistic. Also the nature of the coding made it impossible to simulate the detailed phonetic differences in Marcus et al.'s (1999) Experiment 1 versus Experiments 2 and 3.

At a minimum, the foregoing models provide existence proofs that connectionist networks can cover the Marcus et al. (1999) infant data. Moreover, six of the eight models use conventional unstructured networks, showing that symbolic rules with bound variables are not required.

PROPOSED MODEL

Our work here offers a simulation like that of Shultz (1999), but with phonetically realistic encoding of the input sentences.[1] If successful, it would suggest that analog coding could actually be used by infants on this artificial grammar task. The new coding scheme uses a continuous sonority scale. Sonority is the notion of vowel likeness. Our scale capitalizes on three interesting features of sonority: that some vowels are more vowellike than others, that even consonants vary in their similarity to vowels, and that there are semivowels in the gray area between vowels and consonants.

Learning Algorithm

[1] A preliminary report of this simulation was presented in Shultz and Bale (2000).

Following Shultz (1999), we used an encoder version of the cascade-correlation learning algorithm. Cascade-correlation is a generative algorithm for learning from examples in feed-forward neural networks (Fahlman & Lebiere, 1990). As with other generative algorithms, cascade-correlation constructs its own network topology as it learns by recruiting new hidden units as needed, thus effectively searching in topology space as well as in weight space for a solution. *Topology space* is the space of possible network topologies, ordinarily searched by hand by modelers using static networks. *Weight space* is the space of different patterns of network weights, ordinarily searched automatically by a learning algorithm except in the case of hand-designed weights. As noted, unlike the more standard, back-propagation networks with designed and static topologies, cascade-correlation networks grow as well as learn (Fahlman & Lebiere, 1990). They grow during so-called input phases by recruiting new hidden units into the network as required to reduce error. New hidden units are recruited one at a time and installed each on a separate layer, receiving input from the input units and from any existing hidden units. The candidate hidden unit that gets recruited is the one whose activations correlate most highly with the current error of the network. After recruiting a new unit, the network returns to the so-called output phase in which weights feeding the output units are adjusted to reduce error. See Appendix 1 at http://www.infancyarchives.com for further explanation of cascade-correlation and the encoder option.

Some neurological justification for generative networks such as cascade-correlation is provided by recent findings on learning-driven neurogenesis and synaptogenesis throughout the lifespan (Eriksson et al., 1998; Gould, Tanapat, Hastings, & Shors, 1999; Kempermann, Kuhn, & Gage, 1997; Quartz & Sejnowski, 1997). Neurogenesis and neural migration may be too slow to account for learning within the time frame of the typical infant familiarization experiment, but synaptogenesis can occur within seconds (Bolshakov, Golan, Kandel, & Siegelbaum, 1997). Although cascade correlation is in the abstract neurologically plausible, like most cognitive models it does not provide a detailed model of neural circuits. Also, like all other neural network learning algorithms, it uses some mathematical shortcuts for purely neural processes.

Cascade-correlation has been used to simulate many aspects of cognitive development in older children, including the balance scale (Shultz, Mareschal, & Schmidt, 1994); conservation (Shultz, 1998); seriation (Mareschal & Shultz, 1999); pronoun semantics (Oshima-Takane, Takane, & Shultz, 1999); discrimination shift learning (Sirois & Shultz, 1998); and integration of velocity, time, and distance cues (Buckingham & Shultz, 2000). In these models, network behavior becomes rule-like with learning but knowledge is not represented in rules and cognitive processing is not accomplished by explicit variable binding and rule firing. Rules are instead viewed as abstract, epiphenomenal characterizations of processes at the subsymbolic level of unit activations and connection weights (Smolensky, 1988). Among the advantages of implementing rule-like behavior with neural processes are acquisition of psychologically realistic nonnormative rules (Buckingham & Shultz, 2000), integration of perceptual and cognitive phenomena (Mareschal & Shultz, 1999; Shultz, 1998; Shultz et al., 1994), achievement of the right degree of crispness in knowledge representations (Shultz, 1999), and natural variation across problems and individuals (cf. any of the cascade-correlation simulations).

An encoder option to cascade-correlation (Shultz, 1999) freezes direct input-output connections at 0 to prevent trivial solutions in which weights of about 1 are learned between each input unit and its corresponding output unit. Such trivial solutions solve an encoder problem very quickly in the sense of error-free performance but tend not to develop interesting knowledge

representations that could enable completion of partial patterns. Again, with an encoder network the task is construed as learning to recognize words and sentences.

In short, we use cascade-correlation because it learns deeply and quickly with a minimal-sized network that then generalizes well. It has also been used in many of our other simulations of development and the fact that it grows is consistent with neurological findings of synaptogenesis. Furthermore, we don't need to explore topology space to handdesign the networks; the algorithm does this automatically by constructing networks just large enough to solve the task. Finally, cascade-correlation can be used in encoder mode, which makes it suitable for simulating habituation.

Coding Scheme

We employed a coding scheme with a continuous sonority scale inspired by a tradition of work in phonology (Vroomen, van den Bosch, & de Gelder, 1998). *Sonority* can be informally defined as the quality of vowel likeness. More formally, there are three different definitions. Sonority can be defined perceptually through specifications of saliency (Price, 1980). It can be defined via articulation by measuring the openness of the vocal tract (Selkirk, 1984). Or it can be defined as an epiphenomenon of a feature based system (Clements, 1990). Regardless of the choice of definition, the functional effects of sonority are well documented in terms of syllabification (Clements, 1990; Jespersen, 1922; Harris, 1983). The choice of definition has no particular effect on the performance of our simulations.

The coding scheme for the phonemes making up the one-syllable words in the infant experiments is shown in Table 2. The specific numbers used to represent sonority in our coding scheme are somewhat arbitrary, but the hierarchy is based on work by Vroomen et al. (1998), who in turn based their sonority hierarchy on that of Selkirk (1984).

Table 2

Phoneme Sonority Scale

Phoneme category	Examples	Sonority
low vowels	/a/ /æ/	6
mid vowels	/ɛ/ /e/ /o/ /ɔ/	5
high vowels	/I/ /i/ /U/ /u/	4
semi-vowels and laterals	/w/ /y/ /l/	-1
nasals	/n/ /m/ / ŋ /	-2
voiced fricatives	/z/ /ʒ/ /v/	-3
voiceless fricatives	/s/ /ʃ/ /f/	-4
voiced stops	/b/ /d/ /g/	-5
voiceless stops	/p/ /t/ /k/	-6

Note. Example phonemes are represented in International Phonetic Alphabet. From "Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables." By T. R. Shultz and A. C. Bale. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (p. 461), 2000. Mahwah, NJ: Erlbaum. Copyright 2000 by the Cognitive Science Society, Inc. Adapted by permission.

As seen in Table 2, our sonorities range from -6 to 6 in steps of 1, with a substantial gap and change of sign between the major categories of consonants and vowels. Syllabified consonants (syllabic L, N, and R) were not used in the Marcus et al. (1999) experiments. It is assumed that these syllabified consonants have a sonority value that bridge the consonant-vowel distinction; that is, they have a sonority value between semivowels and high vowels. Each word was coded on two units for the sonority of its consonant and the sonority of its vowel. This is similar to Negishi's (1999b) coding except that consonants and vowels are placed on a single scale rather than on separate scales. In that sense, our coding scheme is a bit more parsimonious than Negishi's (1999b). We coded each sentence in the artificial language with six units, two for each one-syllable word. For example, the sentence *ga ti ga* was coded as (-5 6 -6 4 -5 6).

Our coding scheme for phonemes is deliberately simple and may eventually need expanding. For example, using a one-dimensional sonority scale entails that no coding distinctions are made between phonemes of the same sonority. An example of overlap would be *ba* (an A word) and *ga* (a B word), both having a sonority representation of (-5 6). Sonority values for all the words used in our simulations of the three experiments are provided in Table 3.

Table 3

Sonority Values for Words used in Simulations of Experiments 1-3

Experiment	Category	Word	Consonant	Vowel
1	Train A[a]	*ga*	-5.0	6.0
		li	-1.0	4.0
		ni	-2.0	4.0
		ta	-6.0	6.0
	Train B[a]	*ti*	-6.0	4.0
		na	-2.0	6.0
		gi	-5.0	4.0
		la	-1.0	6.0
	Test A[b]	*wo*	-1.0	5.0
		de	-5.0	5.0
	Test B[b]	*fe*	-4.0	5.0
		ko	-6.0	5.0
2 and 3	Train A[a]	*le*	-1.0	5.0
		wi	-1.0	4.0
		ji	-3.0	4.0
		de	-5.0	5.0
	Train B[a]	*di*	-5.0	4.0
		je	-3.0	5.0
		li	-1.0	4.0
		we	-1.0	5.0
	Test A[b]	*ba*	-5.0	6.0
		ko	-6.0	5.0
	Test B[b]	*po*	-6.0	5.0
		ga	-5.0	6.0

[a]All four training words from the A category were combined with all four training words from the B category to create 16 training sentences in each condition of each experiment. [b]The first of

the A test words was combined with the first of the B test words to form one test sentence; the second of the A test words was combined with the second of the B test words to form another test sentence in each condition of each experiment.

Parameters

All cascade-correlation parameters were Fahlman's (1991) default values with these exceptions. *Score threshold*, the tolerated difference between target and actual outputs, was raised from 0.4 to 0.8 to reduce the crispness of the knowledge representations. Training continued until all output units produced activations within score threshold of their targets. The parameters for input patience and output patience were set to 1 rather than the default of 8. As noted earlier, cascade-correlation alternates between two phases – output phase and input phase. During output phases, connection weights entering output units are adjusted to reduce error. During input phases, connection weights entering candidate hidden units are adjusted to increase the correlation between network error and activations on the candidates. The two patience parameters represent the number of epochs allowed to pass with little increase in error reduction or correlation, respectively, before shifting phase. An *epoch* is a presentation of all the training patterns. Patience was reduced because performance did not improve much after it failed to improve on a single epoch. Additional computational and mathematical details about cascade-correlation can be found in Appendix 1 (http://www.infancyarchives.com) and in previous papers (e.g., Fahlman & Lebiere, 1990; Shultz et al., 1994).

Eight networks were run in each condition of the three experiments. Each network, starting with its own randomly determined connection weights, including those initial weights used for candidate hidden units, corresponds to a unique infant in the Marcus et al. (1999) experiments. Output units had linear activation functions to enable their approximation of real numbers; hidden units had sigmoid activation functions.

Experimental Design

All sentences and experimental designs were identical to those used with the Marcus et al. (1999) infants (see Table 1). In the simulation of Experiment 1, networks were familiarized to ABA sentences and then tested on sentences with novel words that were either consistent (ABA) or inconsistent (ABB). In another condition, familiarization was to ABB sentences, with novel ABB sentences as consistent and novel ABA sentences as inconsistent. Experiment 2 employed the same syntactic patterns, but the words were chosen more carefully so that phoneme sequences were different in the familiarization and test patterns. Experiment 3 used the same A and B categories of words as did Experiment 2, but employed contrastive syntactic patterns that each duplicated a consecutive word: AAB versus ABB. For each experiment, network error on test patterns was subjected to a mixed, repeated measures ANOVA in which familiarization condition served as a between-network factor and consistency of test pattern served as a repeated measure.

Results

Mean network error on test patterns for simulations of the three experiments is shown in Table 4, along with F and p values for the main effect of consistency of the test pattern. These results indicate more network error to inconsistent test patterns than to consistent test patterns for each experiment. With error representing the need for further cognitive processing, these results mirror the infant results of Marcus et al. (1999).

Table 4

Mean Network Error on Test Patterns for Simulations of Three Experiments

Experiment	Patterns	Consistent	Inconsistent	$F(1, 14)$	$p<$
1	ABA vs. ABB	8.2	14.5	74	.0001
2	ABA vs. ABB	13.1	15.8	26	.0001
3	AAB vs. ABB	12.9	15.3	52	.0001

Note. A and B refer to two different categories of monosyllabic nonsense words. From "Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables." By T. R. Shultz and A. C. Bale. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (p. 462), 2000. Mahwah, NJ: Erlbaum. Copyright 2000 by the Cognitive Science Society, Inc. Adapted by permission.

Mean and standard deviation of required epochs and hidden units recruited are presented for each condition of each experiment in Table 5. It is not clear how to match the epochs that networks require to learn with trials in infant experiments because it is not known how much processing infants do on each trial and how that can be calibrated with network processing. Cascade-correlation generally learns rather fast compared to many neural network algorithms, and this is reflected in the relatively small number of epochs required.

Table 5

Mean and Standard Deviation of Epochs and Hidden Units in Simulations of Three Experiments

Experiment	Condition	Epochs		Hidden units	
		Mean	SD	Mean	SD
1	ABA	52	26	2.6	1.1
	ABB	46	24	2.4	1.1
2	ABA	101	26	4.9	1.1
	ABB	90	29	4.4	1.4
3	AAB	83	18	4.4	0.9
	ABB	91	32	4.8	1.8

A plot of mean error results over output-phase epochs for one representative network is presented in Figure 1. The first three epochs are omitted from this plot for clarity because error started so high, at around 380. There is an approximate exponential decrease in error on the training patterns over time, much like the shape of declining attention in infant familiarization experiments. After complete success with the training patterns, the consistent test patterns likewise show rather little error, but the inconsistent test patterns show considerable error recovery, as in typical infant studies. The points at which hidden units are recruited into the network are marked with diamonds just above the training errors. In cascade-correlation learning, error typically decreases sharply after a new hidden unit is recruited. Because error on input-phase epochs does not change, they are usually excluded from such plots.

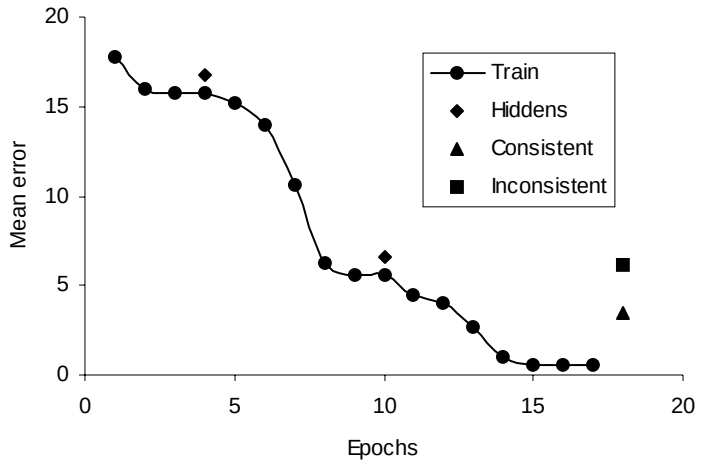
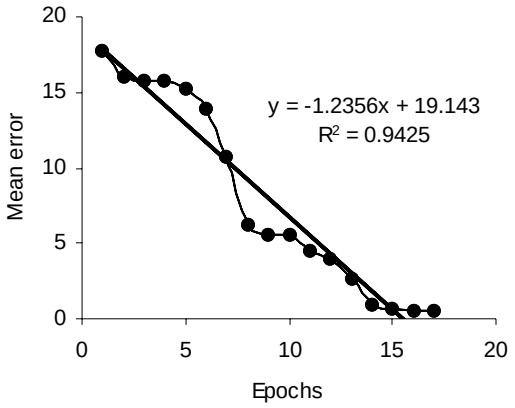


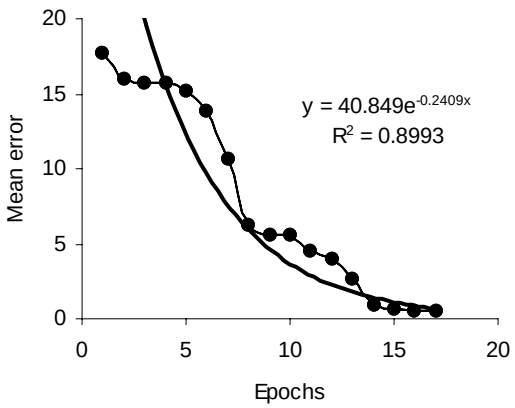
Figure 1. Mean error over consecutive output epochs for a network in the ABA condition of the simulation of Experiment 1.

To investigate the shape of the error reduction curve for this network, the mean training error was fit with linear and exponential functions, both with and without inclusion of the previous (third) epoch. Results of these data fits are plotted in Figure 2, along with the best-fitting function and the R^2 value, or amount of variance accounted for by the best fitting function. Without the third epoch included, the data are fit about equally well with linear (Figure 2a) or exponential (Figure 2b) functions. However, with the previous (third) epoch included, an exponential fit (Figure 2d) is much better than a linear fit (Figure 2c). Inclusion of even earlier epochs (first and second) further worsens a linear fit. Thus, the overall shape of the error reduction is exponential. Other networks produce essentially the same results.

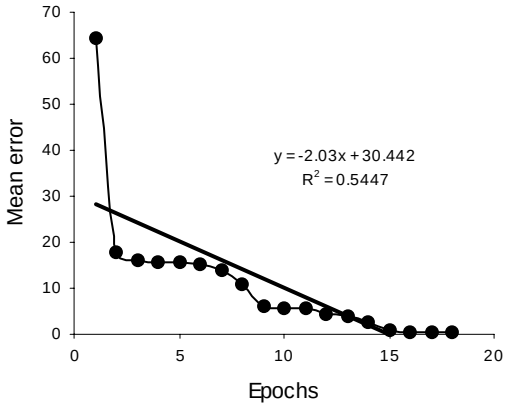
a. Linear fit



b. Exponential fit



c. Linear fit including previous epoch



d. Exponential fit including previous epoch

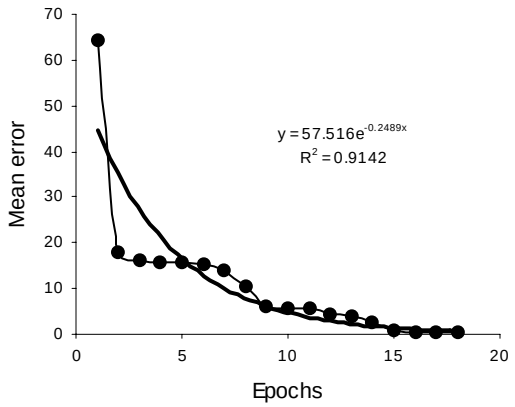


Figure 2. Linear and exponential fits to the training curve of Figure 1, with and without inclusion of the previous epoch.

Discussion

Our network model succeeds in capturing both the basics of Marcus et al.'s (1999) infant experiments (more recovery of attention to inconsistent novel patterns than to consistent novel patterns) and the typical exponential decrease in attention during familiarization. The model captures these phenomena without the explicit rule and variable technique that was claimed by Marcus et al. to be required. As with other unstructured neural models, this shows that at least some of the functionality of symbolic rules and variable binding can be constructed from subsymbolic processes without having to be assumed and explicitly built in by the modeler.

REVERSALS OF PREFERENCE

An interesting feature of the Marcus et al. (1999) results was that a small proportion of infants actually showed a reversal of the general preference trend in attending to test patterns. In other words, rather than attending more to inconsistent test patterns, these exceptional infants attended more to consistent patterns. Increasing the score-threshold parameter to 0.8 made network

learning sloppy enough to bring several simulated infants close to a reversal in the simulations just reported. However, none of them actually showed a reversal. The previous simulation using arbitrary analog coding with cascade-correlation encoder networks with a similarly loose score-threshold setting showed a small number of reversals (Shultz, 1999).

To determine whether our model would show similar reversals, we repeated the simulations just reported but with 20 rather than 8 networks in each of the two conditions of each of the three experiments. All parameter settings remained the same. The idea was to provide a better opportunity for these low probability reversals to occur, merely by increasing the number of observations.

Results

The proportions of networks showing a reversal of the attention trend were 2/40, 4/40, and 2/40 in simulations of Experiments 1, 2, and 3, respectively, yielding an overall proportion of .0667 reversals. All these reversals were very slight. This simulated proportion of reversals is extremely close to that obtained by Marcus et al. (1999). They reported reversal proportions of 1/16, 1/16, and 0/16 infants in Experiments 1, 2, and 3, respectively, yielding an overall proportion of .0417.

Discussion

It is not presently clear what produces reversals of preference in infants in habituation experiments. Nonetheless, the simulated reversals found here are interesting in that they suggest that such reversals could be a natural part of the learning of empirical regularities. When exposure to these regularities is limited, as it was in both the psychological experiments and the simulations, exceptions will naturally occur. This is a theoretically parsimonious explanation of reversals because it does not require assumptions of any extraneous processes such as lack of attention or fussiness on the part of the infants. Symbolic rule-based models may have difficulty capturing such reversals in any natural way if their rules are as crisp as the example rule given by Marcus et al. (1999): "The first item X is the same as the third item Y" (p. 79).

COMPARISON OF TRAINING ERROR TO CONSISTENT TEST ERROR

Figure 1 suggests that generalization to the consistent syntactic patterns is not perfect. At least for this single network, error per pattern was lower for training patterns at the end of training than it was for the consistent test patterns. To test whether this trend was statistically reliable, we ran eight networks in each of the two conditions of each experiment, recording the last training error as well as error on the consistent test patterns. We divided each error by the number of patterns on which it was based, 16 training patterns and two consistent test patterns.

Results

In each of the three experiments, the difference between training error and consistent test error was found to be statistically significant, $p < .001$. For example, for Experiment 1, the mean per-pattern error was 0.48 for training patterns and 3.95 for consistent test patterns, $F(1, 14) = 113.68$, $p < .001$.

Further experimentation revealed that manipulating how deeply the networks are allowed to learn modulates the size of this effect. For example, allowing networks to learn for only 30 epochs in Experiment 1 produced per-pattern error means of 6.01 for training patterns and 6.75 for consistent test patterns, $F(1, 14) = 2.95$, $p = .108$, a non-significant difference. In this case,

there was still a reliable difference between consistent and inconsistent test error (M = 10.32), $F(1, 14) = 36.89$, $p < .001$.

Discussion

The fact that networks do not generalize perfectly to consistent test patterns underscores an important difference between unstructured neural networks and symbolic rules with variables. A symbolic rule with variables would presumably predict perfect generalization to consistent test patterns even with novel tokens. That is, the symbolic approach would predict that per-pattern error on consistent test patterns would be as low as per-pattern error on training patterns. If a rule is truly abstract and contains symbolic variables, then it should apply equally well regardless of the particular tokens that the variable is bound to. In contrast, our unstructured neural networks, because of their continuous representations and approximate computation, notice the difference created by novel tokens and show more error to them, even in a familiar pattern, than to the tokens used in training.

Many habituation experiments, including that of Marcus et al. (1999), because of procedure changes between habituation and test phases, do not afford an unambiguous comparison between performance on habituation and test items. However, a cursory examination of a number of habituation experiments in the literature that appeared to retain the same procedure throughout habituation and testing revealed considerable variability on the comparison of interest. In some cases, it appears that generalization performance can be as good as training performance, whereas in other cases it appears that even the test items with the best generalization performance elicit more attention than the habituated items. Based on our simulations, it is reasonable to assume that this variability reflects differences in depth of learning and similarity between training and test items. In any case, our model predicts more attention to syntactically consistent sentences with novel words than to habituation sentences, at least when the habituation sentences are learned to a sufficient depth.

EXTRAPOLATION OUTSIDE THE TRAINING RANGE

Neural networks must generalize well to be taken seriously as cognitive models. If they merely memorize relations between input and output patterns, which they may tend to do if provided with too much computational power (too many hidden units), then they tend not to generalize very well. On the other hand, if neural networks have insufficient computational power (too few hidden units), then they may not be able to learn a particular task. The trick then in designing static neural networks is to predict the minimum number of hidden units required to learn a task. A lean but sufficiently powerful network will typically generalize well to patterns not used in training.

Cascade-correlation removes some of the uncertainty and mystery about network design by starting networks with no hidden units and gradually building up computational power as needed to learn a problem. This ensures that networks have only as much power as needed. Coverage of the Marcus et al. (1999) data on preference for inconsistent over consistent novel patterns shows that cascade-correlation networks indeed generalize well within the range of the training patterns. Such generalization within the training range is known as *interpolation*. Although neural networks are often good at interpolation, their ability to extrapolate outside the training range has sometimes been questioned (Marcus, 1998; Pinker, 1997).

Method

To determine the extrapolation capacity of our networks, we repeated the simulation of Marcus et al.'s (1999) Experiment 1 with a different set of test patterns. The new test patterns for the A- and B-category words are shown in Table 6. Recall that consonants used in training ranged from -6 to -1 in sonority and that vowels used in training ranged from 4 to 6 in sonority. Test patterns for this simulation were either inside the training range (by a distance of 0.5 from the extreme values), outside but close to the extremes of the training range (by a distance of 0.5 from the extreme values), or farther outside the extremes of the training range (by a distance of 1.0 from the extreme values).

Table 6

Test Patterns for Evaluating Extrapolation in the Simulation of Experiment 1

Distance from training range	Category A		Category B	
	Consonant	Vowel	Consonant	Vowel
Inside	-5.5	5.5	-1.5	4.5
Close	-6.5	6.5	-0.5	3.5
Far	-7.0	7.0	0.0	3.0

Eight networks were familiarized to ABA sentences and eight others were familiarized to ABB sentences as in our original simulation of Experiment 1. All parameters remained the same.

Results

Error on the test patterns after familiarization was subjected to a repeated measures ANOVA in which familiarization condition (ABA vs. ABB) served as a between-network factor and consistency (consistent vs. inconsistent) and distance (inside, close, far) served as repeated measures. There were main effects of consistency, $F(1, 14) = 80$, $p < .001$, and distance $F(2, 28) = 527$, $p < .001$, and an interaction between them, $F(2, 28) = 32$, $p < .001$. Mean error for the consistency by distance interaction is shown in Figure 3. The simple main effect of consistency was significant ($p < .001$) at each level of distance.

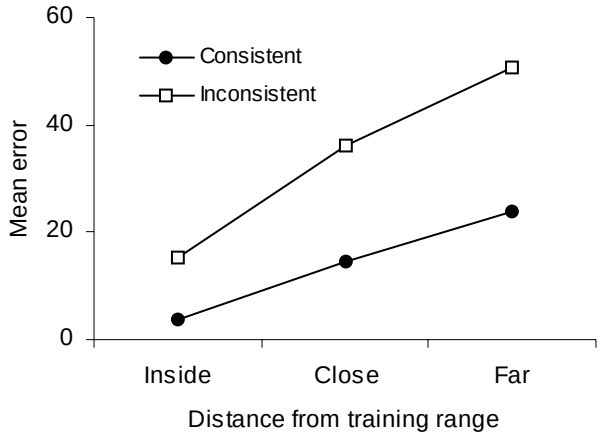


Figure 3. Mean error on consistent and inconsistent test patterns at various distances from the training range for networks familiarized to ABA sentences.

Discussion

Error increased with distance from the training patterns, as expected. The networks generalized better within the range that they were trained on. Also as expected, inconsistent patterns generated more error than did consistent patterns. What is perhaps more counterintuitive is that the consistency effect grew a bit larger with increasing distance from the training patterns. Thus, there is substantial extrapolation ability in these networks.

Moreover, these extrapolation results would appear to generate predictions that stand in sharp contrast to those expected from a rule-based approach. In contrast to the two main effects and interaction generated by our networks, a rule-based approach would presumably predict a single main effect for consistency. A universally quantified rule would mark the distinction between consistent and inconsistent patterns in a manner that disregarded any main or interactive effects of distance from the training range. Thus, if stimuli outside of the training range could be found, an appropriately designed infant experiment on extrapolation outside the habituation range could constitute a critical test of rule-based and connectionist theories.

KNOWLEDGE REPRESENTATION ANALYSIS

It is important to understand how successful neural network models represent their knowledge of the task they are learning. This can provide insights into how the networks solve a problem, suggestions for how the simulated participants may be representing this problem, and predictions for future simulations or psychological study. Detailed examination of knowledge representations and problem-solving strategies is considerably easier with artificial neural networks than with children, although even networks can be somewhat challenging to analyze.

In general, networks can be analyzed by examining unit activations, connection weights, or contributions, which are products of sending-unit activations and connection weights. Patterns of unit activations represent momentary changes in the active memory of a network, whereas patterns of connection weights represent the network's long-term knowledge of a problem. Activations on output units reveal network responses to particular stimuli, whereas activations on hidden units reveal active-memory network knowledge representations of particular stimuli. Contributions are products of sending-unit activations and connection weights entering output units (Sanger, 1989), thus taking account of both active and long-term knowledge as it is summarized at the level of network output.

Because these knowledge representation analyses are lengthy and technical, they are stored in Appendix 2 at http://www.infancyarchives.com. Results of the knowledge representation analyses can be summarized as follows:

1. Networks learn to encode syllables as a linear combination of consonant and vowel sonority.
2. Because networks try to reduce as much error as possible, the first recruited hidden unit learns to encode the duplicate word and the second recruited hidden unit learns to encode the single word in each three-word sentence.
3. Bias weights learn to encode the distinction between consonants and vowels.
4. Networks learn to decode duplicate words with very similar sets of weights to the output units that represent the duplicate words, although this simple strategy fails when the network is exposed to more than one syntactic pattern. With multiple syntactic patterns, networks are forced to learn more complex representations, at least some of which can be identified by contribution analysis.

LEARNING DUPLICATE VERSUS SINGLE WORDS

The knowledge representation analyses of networks learning one sentence pattern predict that these networks would learn duplicate-word categories before single-word categories. Connection weights to outputs representing duplicate words grow large before connection weights to outputs representing single words. Components representing variation in sonority of duplicate words appear in principle component analyses of contributions before components representing variation in sonority of single words. finally, activity in the first hidden unit reflects variation in sonority of the duplicate-word category, whereas activity in the second hidden unit reflects variation in sonority of the single-word category. This simulation tests the prediction that networks learn to recognize the duplicate word before the single word in these three-word sentences.

Method

To test this prediction within the model, we ran a few networks in each condition of each experiment while recording network error for each output unit at the end of each output phase and at the end of training. Mean error was computed for the A words and B words separately.

Results

Figure 4 plots the mean error to A- and B-category words in two networks, one trained on ABA sentences and the other on ABB sentences, after recruiting one and two hidden units. These particular networks are from Experiment 1. After adjusting to one hidden unit, the network trained on ABA sentences had mastered the duplicate A words, but still showed considerable error on the B words. Similarly, the network trained on ABB sentences, after adjusting to one hidden unit, had mastered the duplicate B words, but not the A words. After training, with two hidden units, both networks showed the expected mastery of both word categories.

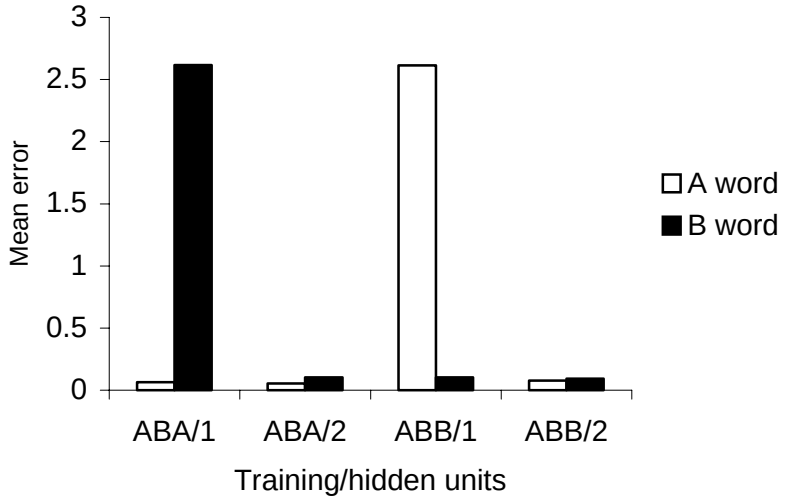


Figure 4. Mean error to A- and B-category words in two networks, one trained on ABA sentences and the other on ABB sentences from Experiment 1, after recruiting one and two hidden units. This confirms that networks learn duplicated words before single words.

Discussion

This pattern of results, which is representative of all of the networks that we examined, verifies the prediction that networks learn the duplicate-word category before the single-word category within whatever sentence type they are exposed to. This is natural for artificial neural networks because although their learning tries to deal with all error at once, the largest error has the largest effect on weight adjustments. As such, this pattern of results would very likely hold for conventional static network models as well as the current generative networks. The tendency to learn duplicate-word categories before single-word categories may be somewhat difficult to test with infants, but appears to be a strong, natural prediction of neural network models.

NATURE OF NETWORK ERROR ON TEST PATTERNS

Given that duplicate words tend to be abstracted into a single representation, it is reasonable to ask whether networks process a whole sentence or simply process the single word and one instance of the duplicate word. To answer this and also to understand the nature of network error on test patterns better, we ran eight networks in each condition of the simulations of Experiments 1 and 3. Activations produced on output units when processing test patterns were recorded after the familiarization phase.

Results

Plots of target and actual sonority sums for three syllables on two kinds of test sentences from a representative network familiarized to the ABB pattern in the simulation of Experiment 3 are shown in Figure 5. Discrepancy between actual sonority sums and the corresponding target sonority sums enables visual detection of the principal sources of error.

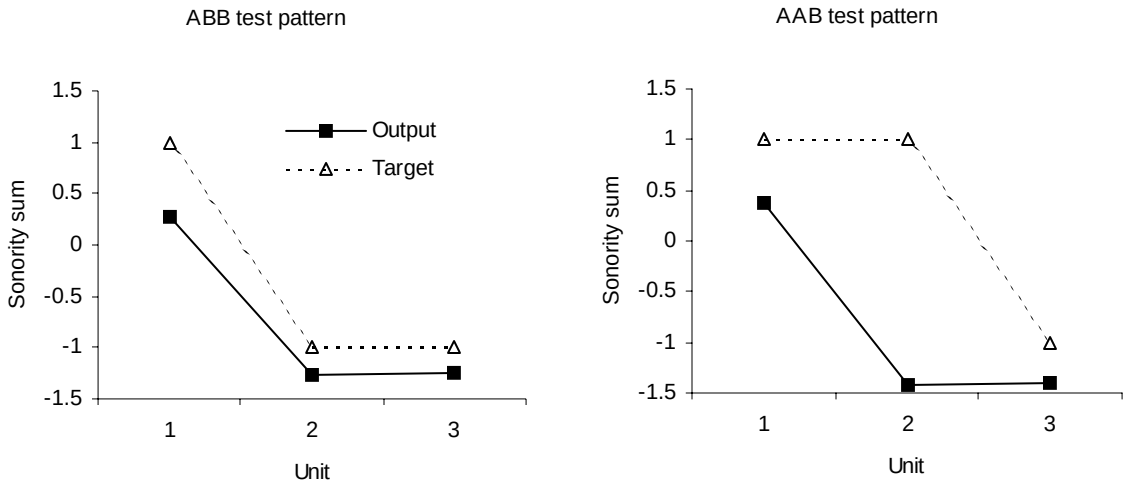


Figure 5. Target and actual sonority sums on two kinds of test sentences from a representative network familiarized to the ABB pattern in the simulation of Experiment 3.

For the network whose results are shown in Figure 5, actual sonority sums on the consistent, ABB test patterns are quite close to their ABB targets. In this particular test sentence, the sonority sum of the A word exceeds that of the B word. When tested on the inconsistent, AAB test sentence, there is a sizeable discrepancy on the second syllable. It is as though the network expected an ABB pattern, but encountered an AAB pattern and detected the difference. In this experiment then, most of the test error is associated with the second word in each inconsistent three-word sentence. In contrast, for the simulation of Experiment 1, most of the

error is associated with the third word of inconsistent test sentences because the contrast in test patterns is between ABA and ABB, which differ only on the third word.

Discussion

All the 16 networks tested in each of these two experiments show similar error patterns. That is to say that most of the error a network arises on the syllable that is unexpectedly different from the familiarized pattern. Networks are essentially correct on every word except the word that deviates from the expected syntax. This pattern underscores that networks are processing the whole sentence even though the duplicate words are largely redundant in the knowledge representations of the networks as revealed by the knowledge representation analysis. It also verifies that error occurs where it should occur if networks recognize a syntactic pattern; that is, error is concentrated on the unexpected word category and positio

PREDICTING AN IMPOSSIBLE DISCRIMINATION: A THOUGHT SIMULATION

Our model can be used to design sentences for which there would be no difference between consistent and inconsistent test patterns, essentially because it would be impossible to distinguish the A and B categories. To illustrate this, consider the actual test sentences that were used, presented with their consonant and vowel sonority values in Table 7. Because of the overlap in sonority values, it is somewhat remarkable that networks were able to distinguish between ABA and ABB sentences. The words *ba* and *ga* have identical sonority values as do the words *po* and *ko*. Despite this overlap, each test sentence is unique because of the particular interactions between syntactic patterns and categories.

Table 7

Test Sentences and Sonorities used in Experiment 2

Pattern	Word 1	Word 2	Word 3
ABA 1	*ba* (-5 6)	*po* (-6 5)	*ba* (-5 6)
ABA 2	*ko* (-6 5)	*ga* (-5 6)	*ko* (-6 5)
ABB 1	*ba* (-5 6)	*po* (-6 5)	*po* (-6 5)
ABB 2	*ko* (-6 5)	*ga* (-5 6)	*ga* (-5 6)

Note. A and B refer to two different categories of monosyllabic nonsense words.

We can make the task much more difficult merely by interchanging the words *po* and *ga* in these sentences, as illustrated in Table 8. With this change, the first ABA sentence is identical to the first ABB sentence, and the second ABA sentence is identical to the second ABB sentence, in terms of sonority. Thus, there would be no way for networks to distinguish consistent from inconsistent sentences on sonority alone. We call this a *thought simulation*, because there is no need to actually run it. As with a thought experiment, just thinking about this hypothetical simulation is sufficient to see that the consistent and inconsistent tests would be indistinguishable to the networks. In this case, the sonority coding provides insufficient information to distinguish the two syntactic test patterns. It would, of course, be possible to design training sentences with these same characteristics, in which case the two training conditions (e.g., ABA vs. ABB) would produce indistinguishable test results.

Table 8

Some Difficult-to-discriminate Test Sentences and Sonorities Created by Interchanging the Words *po* and *ga* in Table 7

Pattern	Word 1	Word 2	Word 3
ABA 1	*ba* (-5 6)	*ga* (-5 6)	*ba* (-5 6)
ABA 2	*ko* (-6 5)	*po* (-6 5)	*ko* (-6 5)
ABB 1	*ba* (-5 6)	*ga* (-5 6)	*ga* (-5 6)
ABB 2	*ko* (-6 5)	*po* (-6 5)	*po* (-6 5)

Note. A and B refer to two different categories of monosyllabic nonsense words.

If infants in similar experiments produced similar results, this would provide strong support for our model. In contrast, if infants showed as much recovery to inconsistent patterns as they did in the original Marcus et al. (1999) experiments, then our coding scheme would be called into question. If the infant results were somewhere in between these two possibilities, that is, less of a difference but still a difference, it might be reasonable to make our sonority scale a bit more discriminating. This may be done by adding phonetically motivated scale levels to distinguish, for example, *ba* from *ga* and *ko* from *po*. Simulations with a more differentiated scale would probably show an intermediate level of discrimination between the sentence types, just as the hypothetical infants may.

Another possibility is that phonemes could be coded with binary phonetic features as well as sonority. This could happen either in parallel (at the input level) or in serial, with sonority values being computed from binary phonetic features. Such additional binary information about phonemes may be exploited to distinguish between syntactic categories and thus syntactic patterns. Note that this prediction concerns the relative difficulty of discriminating syntactic patterns of sentences, not perception of phonemes.

GENERAL DISCUSSION

As with several alternative simulations, these results show that an unstructured neural network model without symbolic rules can simulate infant familiarization and novelty results. Like infants, our networks gradually ignored a repeated syntactic form and recovered interest to an inconsistent novel form but not to a consistent novel form, assuming that network error corresponds to a need for further processing. Thus, the infant results of Marcus et al. (1999) do not uniquely require a symbolic rule-based account. The argument that these infant data suggest an innate rule-learning capacity (Pinker, 1999) is at best premature.

Including our present study, there are now at least nine different models of the Marcus et al. (1999) data set, ironically making it one of the most modeled in psychology. All of these models are connectionist, and only two of the nine use explicit variable binding (Gasser & Colunga, 1999); Shastri & Chang, 1999). None of these models use explicit rules along with explicit variable binding, as would be characteristic of symbolic production systems (Anderson & Lebiere, 1998; Newell, 1990). There is consequently no question that the initial claims of Marcus et al. (1999) were overstated and that it is possible for even unstructured neural networks to capture these data. Explicit rules and variables are clearly not needed for a coherent theoretical account of the infant data.

Assuming that issue to be settled, it becomes of greater importance to consider which of the current models offers the best account of these data given that the models are all somewhat different. In contrast to alternative models of the Marcus et al. (1999) data, our model has several features to recommend it. Our model does not require extensive preexperiment experience (Christiansen & Curtin, 1999; Elman, 1999), extensive hand-wiring of networks (Gasser & Colunga, 1999; Shastri & Chang, 1999), external feedback signals not available in the stimuli (Elman, 1999; Shastri & Chang, 1999), unusual interpretation of outputs (Altmann & Dienes, 1999), or explicit variable binding (Shastri & Chang, 1999). On grounds of theoretical parsimony, the more unsupported assumptions that a model requires the less plausible it becomes. A reasonable criterion for assessing competing models is the ease and naturalness with which a data set can be covered.

Unlike some alternative models (Shastri & Chang, 1999; Shultz, 1999; Sirois et al., 2000), our model uses realistic coding of the stimuli. Like Negishi's (1999b) model, ours uses an analog coding of inputs based on the manner in which the phonemes are produced. As noted, our representation of the sentences is a bit more compact and uniform than Negishi's (1999b) because we use a single sonority scale for both consonants and vowels, whereas he used two separate scales, one for place of consonant articulation and another for vowel height. Although some models work fine on familiarization to a single sentence pattern such as ABA, they would likely have difficulty learning two patterns simultaneously, such as ABA and ABB (Negishi, 1999b; Sirois et al., 2000). Such difficulty would stem from a lack of nonlinear hidden units. Learning simultaneous, nonorthogonal sentence patterns constitutes a realistic extension of the Marcus et al. (1999) experiments and was shown to be feasible within our present model.

Our model and those of Shultz (1999) and Sirois et al. (2000) are the only ones so far to capture the other established feature of the Marcus et al. (1999) infant data: the occasional reversal of preference for novel patterns. Just over four percent of their infants and six percent of our networks showed a preference for inconsistent patterns. It is difficult to comment on how easily the other six models may be able to capture these reversals in some natural fashion. However, there are hints that it may be relatively difficult for some models. The Elman (1999) model, for example, showed such a strong consistency effect that occasional reversals of preference seem unlikely: mean activation to ABB sentences was 123 times higher than to ABA sentences. Similarly, the Shastri and Chang (1999) model learns such a powerful representation of serial position of words that reversals would be unlikely to occur. The correlation between weights to position nodes were .9993 for positions 1 and 3 in networks habituated to ABA sentences and .9998 for positions 2 and 3 in networks habituated to ABB sentences. This rather crisp representation resulted, for example, in 3.4 times more error to inconsistent than to consistent sentences in the ABA condition of Experiment 1. Although the mean error differences in Negishi's (1999b) model were more in line with ours and with those of the infants tested by Marcus et al. (1999), no reversals of preference were reported. All of this is not to claim that these alternative models are incapable of matching the proportion of preference reversals found with infants, just that they do not presently do so.

In summary, our model may be currently preferred because it covers more of the infant data, with less preexperimental experience, less network design, and more realistic stimulus coding than alternative models. It also uses a general learning algorithm that has been applied successfully to several other phenomena in cognitive development. Of course, things can change quickly in the modeling literature, and an even better model may be just around the corner.

The power of our model (and those of Shultz, 1999 and Sirois et al., 2000) to cover the infant data probably derives from construing the task as one of recognizing whole sentences. The task of predicting the next word in a sentence, used in several of the alternative neural models, is a difficult one and not necessarily the primary task that humans face when listening to language. Because there is successful coverage of the basic consistency effect even with binary coding, with both cascade-correlation (Shultz, 1999) and auto-associator (Sirois et al., 2000) algorithms, it is clear that analog coding is not absolutely required. However, we prefer to use analog coding, in this and some other simulation domains, because it typically makes for smaller, easier to understand networks and strong generalization.

On the assumption that reduction of network error represents declining need for attention and processing, plots of error reduction over time confirmed that our networks performed like the infants that they were designed to simulate. There was an exponential decrease in error followed by more recovery to inconsistent than to consistent novel patterns. It is not clear how symbolic rule-based models would be able to capture this pattern of data because many rule-based algorithms learn their rules in a single trial when presented with appropriate information (e.g., Anderson & Lebiere, 1998; Newell, 1990). However, for familiarization experiments such learning would have to be gradual with decreasing exponential shape. The model of Shultz (1999) also showed this exponential shape. Whether the other seven alternative connectionist models would show this pattern of habituation remains to be seen.

Unlike hypothetical, symbolic rule-based models, our model shows more error to (interest in) consistent novel patterns than training patterns. Because this effect reflects the continuous representations acquired in unstructured neural network learning, it would probably occur in other neural network models as well. It is as if the networks notice a change in tokens used within a familiar syntactic structure and respond with correspondingly greater interest.

Our networks generalized not only to novel sentences within the range of the training patterns but also to novel sentences outside of the training range. That is, the networks not only interpolated, they also extrapolated. Contrary to some expressed misgivings (e.g., Marcus, 1998; Pinker, 1997), this shows that neural networks are not merely memorizing associations between input and output patterns but are abstracting functions relating inputs to outputs. These abstract functions enable simulation of understanding and generalization by neural networks. It is true that networks learn to associate output patterns with input patterns, but in doing so they also build abstract functions capable of converting inputs to outputs. This kind of abstraction occurs as long as the networks are somewhat underpowered, which is typically the case for generative networks such as cascade-correlation.

It was argued that neural networks may be able to learn abstractions on hidden units (McClelland & Plaut, 1999) that would enable generalization to the novel patterns of the Marcus et al. (1999) experiments. Our extrapolation simulations coupled with our knowledge representation analyses demonstrate that this is indeed the case. Networks generalized outside the range of stimuli seen in training by developing internal representations of sonority variation in two syntactic categories of words. These internal representations allowed networks to recognize syntactic differences in sentences containing words with sonorities outside the training range. The test sentences used by Marcus et al. in fact fell within the sonority range of their training sentences, although with novel combinations of sonority values not used in training (see Table 3). Our extrapolation simulation provides evidence of even more impressive generalization

than seen in the Marcus et al. infants because the simulated test sonority values themselves, not only their consonant-vowel combinations, were novel (see Table 6).

The extrapolation pattern was for error to increase with distance from the training range and with inconsistency of the test patterns and for the consistency effect to increase with distance. Because a rule-based approach predicts a consistency effect unaffected by distance, it may be possible to design a critical experiment with infants whose results could distinguish between the two theoretical approaches.

The fact that error increases with distance from the training range suggests that words are not compared by an equality operator (which would have the same error regardless of the amount of the input). Instead, networks compare words with a similarity operator that is sensitive to the quantities of sonority values. This represents a major difference between symbolic and sub-symbolic accounts.

Analysis of the knowledge representations acquired by our cascade-correlation networks during the familiarization phase revealed the nature of the function that was abstracted. Hidden units used sonority variation in the phonemes of the sentence to represent the sonority variation first of the duplicated-word category and second of the single-word category. This representation is more compact and abstract than that in the inputs because it utilizes the redundancy created by the duplicated word. The network then decodes this hidden unit representation by learning output weights that reproduce the sonority values of the sentence onto output-unit activations. This decoding task is simplified by using very similar weights to outputs representing the duplicate-word category. However, extensions to networks learning to recognize two syntactic forms simultaneously (e.g., ABA and ABB) indicate that networks do not always rely on simple weight-duplication strategies.

In several respects, our knowledge representation analyses were clarified by using sonority sums (vowel sonority plus consonant sonority), or equivalently, sonority differences (vowel sonority minus consonant sonority). As compared to raw sonorities, using sonority sums or differences enabled better separation of component scores of sentences in the PCA of network contributions and stronger linear functions relating sonority to hidden unit activation. No more direct evidence of network summing or subtracting sonorities could be found, although such processes would not be out of the question when it is recalled that network units sum their weighted inputs.

It is interesting that networks learn to parse syllables by grouping together two phonemes into one representation of a word or syllable. In this way a network seems to form syllables from phonemes. Although the use of sonority sums has little justification within the psychological literature, the use of sonority differences does. The difference between the onset consonant and the vowel of a syllable can be interpreted as a calculation of sonority slope (or change). Assuming an arbitrary distance of 1 unit, vowel sonority minus the consonant sonority equals sonority slope. It has been argued that sonority slope can be used to detect syllable boundaries or word boundaries (Vroomen et al., 1998). Sonority differences and sonority sums are essentially similar calculations in that they are both linear operations.

It is sometimes possible to relate network knowledge representations to demonstrated or predicted representations in humans (Shultz, 1998; Sirois & Shultz, 1998). Perhaps the knowledge representations discovered in our networks could be used to guide the study of knowledge representations in infants.

Despite the fact that duplicated words were efficiently represented once on a single dimension, it was also the case that networks processed the entire three-word sentence. The relevant evidence comes from the finding that error was concentrated in the single, duplicated word that was inconsistent with the familiarized pattern. There was very little error on the other two words of the sentence. Because the position of this error-prone word varies with condition and test pattern, each of the three words of each sentence was processed.

Because there are at least seven unstructured neural network methods that cover the Marcus et al. (1999) data, there is no demonstrated need for explicit rules and variables in accounting for these data. Whatever functional power is required can be built through subsymbolic learning. It may be the case that some computation in humans is based on explicit symbolic rules, but the Marcus et al. data do not prove this claim as it applies to infants hearing artificial grammars.

It would be a mistake to conclude that the unstructured connectionist models of the Marcus et al. (1999) data show that infant cognition and language are not rule like. To the contrary, these unstructured connectionist models continue to show that systematic rule-like behavior can be implemented without starting with explicitly formulated symbolic rules and variables and the machinery to process them. Within this framework of unstructured connectionism, computational mechanisms are brain-like units and connections, not symbolic rules and variables. That our network analyses relied on complex statistics and graphics underscores the fact that the networks do not compute by matching, selecting, and firing explicit rules. Merely stating a rule to describe how such a network operates would be a very abstract characterization of its behavior, not a mechanistic explanation of how it functions. In this sense, rules may be more in the heads of psychological theorists than in the heads of their participants, more epiphenomenal than computational. The subsymbolic approach has the advantage of showing how brains may be able to implement rule-like behavior. As well, it generates predictions that can be quite different than those generated by models that begin with symbolically explicit rules and variables. The graduated patterns of dishabituation to syntactically consistent sentences with novel words and of extrapolation outside the training range stand as two examples of these differentiating predictions.

The idea that the Marcus et al. (1999) habituation data can be accounted for by unstructured connectionist models is analogous to other recent critiques of possibly overly rich interpretations of infant habituation data (Bogartz, Shinskey, & Speaker, 1997; Fischer & Bidell, 1991; Haith, 1998). These critics argue that lower-level perceptual interpretations of dishabituation differences must be examined before abstract conceptualizations such as hypotheses and theories and emotional reactions such as surprise can be attributed to young infants. Similarly, we believe that subsymbolic neural accounts must be examined before symbolic rules can be attributed to young infants.

In contrast to alternative models, we have generated a variety of predictions for future psychological research: exponential habituation of attention, more attention to consistent test patterns than to well-habituated patterns, larger consistency effects with increasing distance from the training range, ability to learn multiple syntactic forms simultaneously, relative difficulty distinguishing sentence types when words have similar sonorities, earlier learning of a duplicate-word category than a single-word category, and the use of particular knowledge representations. Such predictions are potentially useful when computational modeling outpaces the original psychological research, as has happened here. Because these initial psychological data are so

easy to capture with a variety of structured and unstructured connectionist techniques, and perhaps symbolic techniques (although that still needs to be demonstrated), it becomes important to collect new infant data to adjudicate among the alternative models. Differential model predictions could play a useful role in guiding future psychological research by suggesting critical experiments to distinguish between models. It is likely that existing and future models will need to adapt to new psychological findings.

It may be interesting to further explore recurrent networks on these tasks. Recurrent networks cycle activation from hidden or output units back into the input layer, thus implementing a sort of working memory for processing temporal sequences (Hertz et al., 1991). Nonrecurrent networks like ours and recurrent networks can both learn to solve problems with a temporal component. The main difference between the two is a trading of space for time in the input coding (Hertz et al.). Recurrent networks process inputs in sequence over time, allowing for sequences of indeterminate length. Nonrecurrent networks represent inputs on different units simultaneously. There is a recurrent version of cascade-correlation that could be tried on sentence familiarization problems. An advantage of recurrent algorithms is their ability to deal with sequential input of indefinite and varying length. Because the Marcus et al. (1999) sentences had a fixed length and were separated by a substantial temporal gap, it is not clear that recurrent networks are required to simulate the resulting phenomena. However, recurrent networks would presumably be required to deal with experiments having continuous speech streams and sentences of indefinite length. It may be interesting to see whether recurrent networks could learn when the task is to recognize a sentence pattern, as in our model, instead of learning to predict the next word or syllable, as has been relatively common with recurrent networks so far.

An issue that our model does not address is the tendency noted in some studies for infants to prefer familiar rather than novel stimuli. Generally, such familiarity preferences occur in younger infants, with more complex stimuli and with shorter exposure times (Hunter & Ames, 1988). Such familiarity preferences have also been found in grammar learning tasks, similar to the one used in the Marcus et al. (1999) experiments (Gomez & Gerken, 1999). It is difficult to see how an error-reduction model such as ours could account for familiarity if it is assumed that interest is a simple linear function of error. As noted, Christiansen and Curtin (1999) obtained a familiarity preference with their back-propagation model of the Marcus et al. data, but the computational basis for this effect remains obscure and it does not fit the Marcus et al. data, which showed primarily a novelty preference. Likewise, the occasional reversals noted in our model do not count as a systematic familiarity preference as has been found in some infant studies. Ultimately, what may be needed to capture systematic familiarity preferences is a model that views interest as a more complex function of error. Until error is brought down to near asymptote, interest may remain relatively high and even increase, producing a familiarity preference. But as error nears asymptote, a novelty preference occurs. For now, familiarity preference remains an open and interesting problem.

Our research can be viewed as part of stream of challenges that have been posed to connectionism and responses to those challenges. Over the past dozen years, there have been a number of claims about what connectionist models cannot do, mainly by researchers committed to the symbolic computational paradigm. A list of such claims periodically grows as challenges are made and then shrinks as challenges are met. In general, this dialog has been healthy for

cognitive science because it has focused connectionists' attention on issues important to others and prodded them to either explore or extend the computational power of neural networks.

ACKNOWLEDGMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This work has benefited from comments from David Buckingham, Jacques Katz, Yuriko Oshima-Takane, Sylvain Sirois, and Yoshio Takane, as well as from the Editor and anonymous reviewers.

REFERENCES

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284*, 875.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Bolshakov, V. Y., Golan, H., Kandel, E. R., & Siegelbaum, S. A. (1997). Recruitment of new sites of synaptic transmission during the cAMP-dependent late phase of LTP at CA3-CA1 synapses in the hippocampus. *Neuron, 19*, 635–651.

Bogartz, R. S., Shinskey, J. L., & Speaker, C. J. (1997). Interpreting infant looking: The event set X event set design. *Developmental Psychology, 33*, 408-422.

Buckingham, D., & Shultz, T. R. (2000). The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development, 1*, 305-345.

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: No need for algebraic rules. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Erlbaum.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckmann (Eds.), *Papers in laboratory phonology 1* (pp. 282-333). Cambridge: Cambridge University Press.

Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19*, 157-180.

Cohen, L. B. (1979). Our developing knowledge of infant perception and cognition. *American Psychologist, 34*, 894-899.

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

Elman, J. L. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al. [Online]. Retrieved April 27, 1999 from the World Wide Web: http//www.crl.ucsd.edu/~elman/Papers/MVRVsim.html

Eriksson, P. S., Perfilieva, E., Bjork- Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A., & Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine, 4*, 1313-1317.

Fahlman, S. E. (1991). Common Lisp implementation of cascade-correlation learning algorithm [Computer program]. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.

Fischer, K. W., & Bidell, T. R. (1991). Constraining nativist inferences about cognitive capacities. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 199-235). Hillsdale, NJ: Erlbaum.

Gasser, M., & Colunga, E. (1999). Babies, variables, and connectionist networks. *Proceedings of the Twenty-first Annual conference of the Cognitive Science Society* (p. 794). Mahwah, NJ: Erlbaum.

Gomez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70,* 109-135.

Gould, E., Tanapat, P., Hastings, N. B., & Shors, T. J. (1999). Neurogenesis in adulthood: A possible role in learning. *Trends in Cognitive Sciences, 3,* 186-192.

Haith, M. M. (1990). Progress in the understanding of sensory and perceptual processes in early infancy. *Merrill-Palmer Quarterly, 36,* 1-26.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development, 21,* 167-179.

Harris, J. (1983). Syllable structure and stress in Spanish: A non-linear analysis. *Linguistic Inquiry Monograph 8*. Cambridge, MA: MIT Press.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison Wesley.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 5, pp. 69-95). Norwood, NJ: Ablex.

Jespersen, O. (1922). *Language, its nature and origin*. New York: Holt.

Kempermann, G., Kuhn, H. G., & Gage, F. H. (1997, April 3). More hippocampal neurons in adult mice living in an enriched environment. *Nature, 386,* 493-495.

Kohonen, T. (1988). *Self-organization and associative memory* (2nd edition). New York: Springer-Verlag.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37,* 243-282.

Marcus, G. F. (1999a, April 16). Do infants learn grammar with algebra or statistics? *Science, 284,* 433.

Marcus, G. F. (1999b, May 7). Response: Rule learning by seven-month-old infants and neural networks. *Science, 284,* 875.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999, January 1). Rule learning by seven-month-old infants. *Science, 283,* 77-80.

Mareschal, D., & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. In *Proceedings of the 19th annual conference of the Cognitive Science Society* (pp. 484-489). Mahwah, NJ: LEA.

Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science, 11*, 149-186.

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3*, 166–168.

Negishi, M. (1999a, April 16). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.

Negishi, M. (1999b). Rule learning by seven-month-old infants and by a simple-recurrent-network [Online]. Retrieved April 16 from the World Wide Web: http://www.cns-web.bu.edu/pub/mnx/sci.html

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5*, 193-207.

Oshima-Takane, Y., Takane, Y., & Shultz, T. R. (1999). The learning of first and second pronouns in English: Network models and analysis. *Journal of Child Language, 26*, 545-575.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Pinker, S. (1999, January 1). Out of the minds of babes. *Science, 283*, 40-41.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21-69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition, 61*, 299-308.

Price, P.J. (1980). Sonority and syllabicity: Acoustic correlates of perception. *Phonetica, 37*, 327-343.

Quartz, S. R, & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences, 20*, 537-596.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. *Advances in Infancy Research, 10*, 1-36.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*, 115-138.

Seidenberg, M. S., & Elman, J. L. (1999, April 16). Do infants learn grammar with algebra or statistics? *Science, 284*, 433.

Selkirk, E. O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds.), *Language sound structure* (pp. 107-136). Cambridge, MA: MIT Press.

Shastri, L. (1999, September 10). Infants learning algebraic rules. *Science, 285*, 1673.

Shastri, L., & Chang, S. (1999). A spatiotemporal connectionist model of algebraic rule-learning. TR-99-011. International Computer Science Institute, Berkeley, CA. [Online]. Retrieved September 10, 1999 from the World Wide Web: www.icsi.berkeley.edu/~shastri/babytalk

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science, 1*, 103-126.

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 665-670). Mahwah, NJ: Erlbaum.

Shultz, T. R., & Bale, A. C. (2000). Infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society* (pp. 459-463). Mahwah, NJ: Erlbaum.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An auto-associator perspective. *Developmental Science, 4*, 442-456.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: Pergamon.

Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllabic structure. *Language and Cognitive Processes, 13*, 193-220.