Infant Familiarization to Artificial Sentences: Rule-like Behavior Without Explicit Rules and Variables

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3A 1B1 Canada

Alan C. Bale (alan_bale@sympatico.ca)
Department of Linguistics; McGill University
Montreal, QC H3A 1G5 Canada

Abstract

A recent study of infant familiarization to artificial sentences claimed to produce data that could only be explained by symbolic rule learning and not by unstructured neural networks. Here we present successful unstructured neural network simulations showing that these data do not uniquely support a rule-based account. In contrast to other neural network simulations, our simulations cover more aspects of the data with fewer assumptions using a more realistic coding scheme based on sonority of phonemes. Our networks show exponential decreases in attention to a repeated sentence pattern, more recovery to novel inconsistent sentences than to novel consistent sentences, some preference reversals, and extrapolation.

One of the most simulated phenomena in developmental psychology is a data set that was claimed to be immune from simulation by unstructured neural networks (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). Although the authors maintained that their results could only be explained by explicit rules and variables, there are now at least eight connectionist simulations of the data, most of which do not use explicit variable binding and none of which use explicit rules. Here we present additional neural simulations of these data, arguing that our model may provide the currently most satisfying account. The paper reviews the relevant infant data, presents various interpretations and models, and then focuses on our current model.

The Infant Data

The relevant experiments familiarized 7-month-old infants to three-word artificial sentences and then tested them 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., *ni la ni*) or an ABB pattern (e.g., *ta gi gi*). There were 16 of these sentences, constructed 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 tested with two sentences having novel words that were either consistent or inconsistent with the familiar pattern.

Table 1: Marcus et al. (1999) experiments.

Pattern	Experiments 1 & 2		Experiment 3	
	Cond. 1	Cond. 2	Cond. 1	Cond. 2
Familiarize	ABA	ABB	ABB	AAB
Consistent	ABA	ABB	ABB	AAB
Inconsistent	ABB	ABA	AAB	ABB

When an infant looked at a flashing light to the left or right, a test sentence was played from a speaker situated next to the light. Each test sentence was played until the infant either looked away or 15 s elapsed. Infants attended more to inconsistent novel sentences than to consistent novel sentences, showing that they distinguished the two sentence types.

Experiment 2 was the same except that the words were chosen more carefully so that phoneme sequences were different in the familiarization and test patterns. Experiment 3 used the same words as Experiment 2, but in contrastive syntactic patterns that each duplicated a consecutive word: AAB vs. ABB. The idea was to rule out the possibility that infants might have used the presence or absence of consecutively duplicated words to distinguish sentence types.

In all three experiments, infants attended more to inconsistent than to consistent novel sentences. Our concern is with the best theoretical account of these data. Is the infant cognition based on rules and variables or on connections?

A Rule and Variable Interpretation

Marcus et al. (1999) argued that these grammars could not be learned by the statistical methods common to standard neural networks. They also tried some unsuccessful neural network simulations using Simple Recurrent Networks (SRN). The authors proposed that a only 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)." They allowed that their data might also be accounted for by 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)."

Psychology of Familiarization

A leading psychological analysis of familiarization assumes that infants build categories for stimuli (Cohen, 1973; Sokolov, 1963). Subsequently, they ignore stimuli that correspond to their categories, and concentrate on stimuli that are relatively novel. These processes are often discussed in terms of recognition memory. If there is substantial recovery to a novel test stimulus, then it is considered novel. But if there is little or no recovery, then the stimulus is considered to be recognized as a member of a familiar category. During familiarization there is typically an exponential decrease in attention.

Familiarization in Neural Networks

Encoder networks that learn to reproduce their inputs on their output units can simulate familiarization and novelty effects in infants (Mareschal & French, 1997). Relations among stimulus features are encoded in hidden unit representations, and accuracy is tested by decoding these hidden unit representations onto output units. Discrepancy between output and input representations is network error. Familiar stimuli produce less error than novel stimuli, which presumably deserve further learning. Such hidden unit representations enable prototypes, generalization, and pattern completion (Hertz, Krogh, & Palmer, 1991).

Other Neural Network Models

There are at least 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 not be able to simulate the data. Most of these models are ordinary unstructured connectionist models without explicit rules and variables. All eight of these models cover the basic finding of the Marcus et al. (1999) experiments, namely noticing the difference between consistent and inconsistent sentences. It is beyond the scope of this brief paper to thoroughly review all of these models, many of which are as yet only sketchily reported. However, we can briefly characterize each model and identify what we believe to be its best virtue and most significant limitation.

Four of the unstructured models use the SRN architecture, construing the network's task to be prediction of the next word in a sentence. Negishi (1999a, b) used an SRN without hidden units, coding each word in analog fashion with place of consonant articulation and vowel height. This is a simple network requiring no unusual hand-wired assumptions or pre-experimental experience. However, Marcus (1999a) claimed that it essentially implemented variables by using continuous values on the input units that are transmitted directly to the outputs, thus arguably disqualifying the model from meeting the challenge that variable binding is required.

Following an argument that Marcus et al.'s (1999) SRNs failed because they lacked normal phonemic experience (Seidenberg & Elman, 1999), Elman (1999) pre-trained an SRN to distinguish whether each word differed or not from the previous word. Each word was coded on 12 binary phonetic features. Although 7-month-olds obviously know something about phonemes and it may be reasonable to include such knowledge in models, it is unlikely that infants receive any target signals about phonemic sameness and difference. More seriously, the network's task in both the pre-training and habituation phases of the simulation was discrimination rather than habituation as it was for the infants.

Christiansen and Curtin (1999) pre-trained an SRN on word segmentation. The network learned to predict the identity and stress of the next phoneme in sentences from information on 11 binary phonological features and the stress and utterance boundaries of individual phonemes. Presented with the Marcus et al. test sentences, the network then showed slightly better prediction of words occurring in inconsistent than those occurring in consistent sentences. Again, the use of prior knowledge seems reasonable. However, it is unclear why the network would perform better on inconsistent sentences, with which it is less familiar, than on consistent sentences whose pattern it has just learned.

Altmann and Dienes (1999) used SRNs with an extra encoding layer between the input and hidden layers. Unlike some models, this one does not require any questionable pre-training and is performing the habituation task. On the negative side, Marcus (1999b) reports that only when somewhat unconventional correlation and distance measures are used can the network discriminate between consistent and inconsistent sentences. It would be more typical to measure error or relative output activation for such networks.

Gasser and Colunga (1999) used a specially-designed network with micro-relation units whose activations correlated with inputs from two different syntactic categories. Hardwired connections caused similar syllables to be synchronized, producing low activations on the micro-relation units, and dissimilar syllables to be desynchronized, producing high activations on the micro-relation units. No pre-training was necessary, but the hardwiring of connection weights is of questionable psychological validity.

Shastri and Chang (1999; Shastri, 1999) designed a structured connectionist model with explicit variable binding, implemented by temporal synchrony of activations on units representing sequential position and other units representing arbitrary binary word features. The network learned to represent an ABA pattern by firing the first position unit synchronously with the third position unit. This network would seem to generalize well to any novel sentences of three words, regardless of the particular features of the words used. But the network is extensively hand-built, and the critically important feedback signals about the position of words in a sentence are psychologically implausible.

None of the foregoing reports of models include evidence on the course of habituation or provide predictions that could be tested with infants.

Shultz (1999) used an encoder version of the cascade-correlation algorithm with arbitrary analog coding of syllables. With an encoder network, the task is construed as word and sentence recognition. Besides covering the consistency effect, these networks learned the training patterns with an exponential decrease in error and showed occasional reversals of preference that were found with the infants. Because the coding was arbitrary, however, it was not possible to simulate the detailed phonetic differences between Marcus et al.'s (1999) Experiments 1 and 2.

Our Model

Here we present a simulation like that of Shultz (1999), but with phonetically realistic encoding of the input sentences using a continuous sonority scale. A successful result would suggest that such coding could be used by infants in their sentence processing. Sonority is the quality of vowel likeness, and can be defined by perceptual salience (Price, 1980) or by openness of the vocal tract (Selkirk, 1984). The coding scheme is shown in Table 2. The specific numbers are somewhat arbitrary, but their ordering is based on phonological work (Selkirk, 1984; Vroomen, van den Bosch, & de Gelder, 1998).

Table 2: Sonority scale with examples in IPA.

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

Sonorities range from -6 to 6 in steps of 1, with a gap and change of sign between the consonants and vowels. Each word was coded on two units for the sonority of its consonant and that of its vowel. This is similar to Negishi's (1999b) coding, except that we place consonants and vowels on a single scale, rather than on separate scales. We coded each sentence in the artificial language with six units, two for each one-syllable word. For example, the sentence *ni la ni* was coded as (-2 4 -1 6 -2 4).

Our learning algorithm, cascade-correlation, grows during learning by recruiting new hidden units into the network as required to reduce error (Fahlman & Lebiere, 1990). Recruited hidden units are installed each on a separate layer, receiving input from the inputs and from existing hidden units. The candidate hidden unit that gets recruited is the one whose activations correlate best with current error. After recruiting a hidden unit, the network returns to the phase in which weights feeding the output units are adjusted to reduce error. 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.

The cascade-correlation algorithm has been used to simulate many other aspects of cognitive development, including the balance scale (Shultz, Mareschal, & Schmidt, 1994), conservation (Shultz, 1998), seriation (Mareschal & Shultz, 1999), discrimination shift learning (Sirois & Shultz, 1998), pronoun semantics (Oshima-Takane, Takane, & Shultz, 1999), and integration of velocity, time, and distance cues (Buckingham & Shultz, in press).

In these models, network behavior becomes rule-like with learning, but knowledge is clearly not represented in rules and cognitive processing is definitely not accomplished by explicit variable binding and rule firing. Instead, rules are viewed as abstract, epi-phenomenal characterizations of processes occurring at the sub-symbolic level of unit activations and connection weights (Smolensky, 1988).

There are several advantages of implementing rule-like behavior with neural processes, including the acquisition of psychologically realistic non-normative rules, integration of perceptual and cognitive phenomena, natural variation across problems and individuals, and achievement of the right degree of crispness in knowledge representations. In many cases, universally quantified rules are too crisp to model knowledge representations in children.

Neurological justification for generative networks such as cascade-correlation is provided by recent findings on learning-driven neurogenesis and synaptogenesis throughout the lifespan (Quartz & Sejnowski, 1997). Although neurogenesis and neural migration may be too slow to account for learning within the time frame of the typical infant familiarization experiment, there is evidence that synaptogenesis can occur within seconds (Bolshakov, Golan, Kandel, & Siegelbaum, 1997).

Like most models of higher cognition, cascade-correlation is not a model of detailed neural circuits. Instead, it is an abstracted and simplified model that is partly inspired by neural principles. Individual units in cascade-correlation networks may correspond roughly to groups of biological neurons, and connection weights may correspond roughly to neural pathways.

Results

Mean network error on test patterns for the three experiments is shown in Table 3. Main effects of consistency were significant at $p < .0001$. The results show more network error to inconsistent test patterns than to consistent test patterns for each experiment. On the assumption that error represents a need for further cognitive processing, these results capture the infant data.

Table 3: Mean error on test patterns.

Expt.	Patterns	Consistent	Inconsistent
1	ABA v. ABB	8.2	14.5
2	ABA v. ABB	13.1	15.8
3	AAB v. ABB	12.9	15.3

The proportion of networks showing a reversal of the consistency effect was .0667, which is close to the .0625 obtained with infants.

A plot of mean error over epochs for a representative network from the ABB condition of Experiment 1 is shown in Figure 1. The first few epochs are omitted for clarity because error started so high, at around 350. Such plots reveal exponential decreases in error on the training patterns over time, similar to the shape of declining attention in infant familiarization. The epochs at which hidden units are installed are shown with diamond shapes just above the training error. As in most cascade-correlation simulations, error decreases sharply after a hidden unit is recruited. After training, error is higher on inconsistent test patterns than on consistent test patterns.

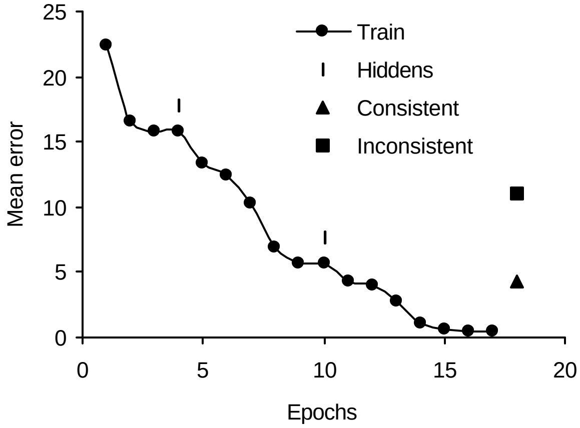

Figure 1: Error reduction in one network.

Generalization tests show that the consistency effect actually grows larger with increasing distance from the training set, a prediction quite different than universally quantified rules would make.

Network analysis revealed that hidden units used sonority sums of consonant and vowel to represent sonority variation first in the duplicated-word category and second in the single-word category. Networks decoded this hidden unit representation with virtually duplicate weights to outputs representing the duplicate-word category.

Discussion

Like other neural models, our model easily captures the consistency effect. In contrast to alternate models of these data, ours has several features to recommend it. Our model does not require extensive pre-experiment 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.

Unlike some alternate models (Shastri & Chang, 1999; Shultz, 1999), our model uses a realistic coding of the stimuli. Like Negishi (1999b), we used an analog coding of inputs based on the manner in which the phonemes are produced. But our representation scheme is a bit more compact and uniform 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. Moreover, our use of hidden units with non-linear transfer functions ensures that any possible variable binding at the input level is lost as activation is propagated forward through the hidden layers.

Our model is the only one so far to capture the other feature of the Marcus et al. (1999) infant data, the occasional reversal of preference for novel patterns. It is unclear how easily other models might be able to capture these reversals, but there are hints that it might be difficult for some models. Elman's (1999) model, for example, had such a strong consistency effect that reversals of preference would be unlikely: mean activation to ABB sentences was 123 times higher than to ABA sentences. Likewise, the Shastri and Chang (1999) model learns a very strong representation of serial position. 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 produced 3.4 times more error to inconsistent than to consistent sentences in the ABA condition of Experiment 1, which would seem to preclude reversals.

Although it is not known why infants show occasional reversals, our simulations show that they can be a natural part learning. With limited exposure, as in both the psychological experiments and our simulations, exceptions naturally occur. This is a parsimonious explanation of reversals because it does not require assumptions of any extraneous processes.

In summary, our model might be currently preferred because it covers more of the infant data, with less pre-experimental experience, less network design, and more realistic stimulus coding than alternate models. It also uses a general learning algorithm that has been applied successfully to several other phenomena in cognitive development.

With so many successful neural models of the consistency effect, there is no question that ordinary, unstructured neural networks can cover these data. The modeling shows that some of the functionality of symbolic rules and variable binding can be constructed from sub-symbolic processes without having to be explicitly built in. The time is now ripe to generate and test predictions from these alternate models.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We thank David Buckingham, Jacques Katz, Yuriko Oshima-Takane, Sylvain Sirois, and Yoshio Takane for comments.

References

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

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.

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

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.

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

Elman, J. L. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al. www.crl.ucsd.edu/~elman/Papers/MVRVsim.html

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.

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.

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

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

Marcus, G. F. (1999b). 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). 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. *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.

Negishi, M. (1999a). 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. www.cns-web.bu.edu/pub/mnx/sci.html

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.

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.

Seidenberg, M. S., & Elman, J. L. (1999). 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). 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. 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., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro, & T. K. Leen, (Eds). *Advances in Neural Information Processing Systems 7* (pp. 601-608). Cambridge, MA: MIT Press.

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*. Hillsdale, NJ: Erlbaum.

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.