Babies, Variables, and Connectionist Networks

Michael Gasser (GASSER@CS.INDIANA.EDU)
Eliana Colunga (ECOLUNGA@CS.INDIANA.EDU)
Computer Science Department, Cognitive Science Program
Indiana University
Bloomington, IN 47405

Abstract

Recent studies have shown that infants have access to what would seem to be highly useful language acquisition skills. On the one hand, they can segment a stream of unmarked syllables into words, based only on the statistical regularities present in it. On the other, they are able to abstract beyond these input-specific regularities and generalize to rules. It has been argued that these are two separate learning mechanisms, that the former is simply associationist whereas the latter requires variables. In this paper we present a neural network model, demonstrating that when a network is made out of the right stuff, specifically, when it has the ability to represent sameness and the ability to represent relations, a simple associationist learning mechanism suffices to perform both of these tasks.

Background

Two recent papers in *Science* have demonstrated the remarkable language learning abilities that are possessed by infants. In both cases the infants were presented with sequences of syllables embodying some sort of regularity and later tested with sequences that agree or disagree in certain ways with the training set. In the experiments of Saffran, Aslin, and Newport (1996), eight-month-olds heard strings of syllables consisting of randomly concatenated three-syllable "words," sequences which never varied internally. Thus the transition probabilities within words were higher than between words. Later the infants were able to differentiate between these words and non-word three-syllable sequences which they had either heard with less frequency than the words or not heard at all. This is taken as evidence that they had picked up the statistics in the training set. Marcus, Vijayan, Rao, and Vishton (1999) presented seven-month-olds with sentences of three one-syllable words separated by gaps. Each sentence sequence consisted of two different words arranged in a fixed pattern, AAB, ABB, or ABA. For example, in the ABB condition, the presented sentences included sequences such as *ga ti ti* and *li na na*. Later the infants responded differently to novel sequences of three words which matched the pattern they had been trained on than to novel sequences which did not. This is taken as evidence that they had in some sense picked up the "rule" implicit in the training sentences.

Marcus et al. (1999) and Pinker (1999) argue that the two studies, taken together, point to at least two distinct learning mechanisms which are behind language learning. One of these, revealed in the experiments of Saffran et al. (1996), can learn relationships such as the tendency for *ti* to immediately follow *ga*. It is sensitive to the content of the items, caring about the similarity among different items. For Pinker (1999), this is just the *associationism* proposed in the eighteenth century by Hume and still proposed as the fundamental mechanism of the mind by modern connectionists and others. The other mechanism, revealed in the experiments of Marcus et al. (1999), can learn relationships such as the fact that the first word of a sentence is the same as the second but different from the third. This mechanism ignores specific content, caring only about sameness or difference. In this sense the second mechanism seems to require *variables*, placeholders which are ignorant of their specific content. For Pinker (1999), this mechanism is an instantiation of what was proposed by the early rationalists and what we think of today as "symbolic." Thus Marcus et al. (1999) and Pinker (1999) now believe that the mind, specifically the portion of it used in language learning, is both associationist and symbolic.

The question, as Marcus et al. (1999) make clear, is not whether neural networks can learn to solve both kinds of tasks, but what sorts of mechanisms are required and whether these differ for the two tasks. In this paper, we will show how a neural network of a particular type (the Playpen architecture) can learn the tasks in Marcus et al.'s experiment, and we will argue that such a network needs at least two mechanisms in addition to those usually found in neural networks, neither of which amounts to explicit variables. Since such a network is also capable of learning the tasks in Saffran et al.'s experiments, we believe that one mechanism suffices to learn the two kinds of knowledge and that that mechanism is associationist.

The Task

Sequences

Before discussing the main issue at hand, the issue of how rule-like behavior could be learned in a neural network, we need to dispense with the more straightforward issue of how patterns in time can be dealt with. The syllable "words" in Marcus et al.'s experiment are presented sequentially, but it is convenient (though perhaps not necessary) to treat the "rule" that is learned as a static pattern. As discussed below, there is a straightforward way to implement this in a neural network using connections with varying time delays. Given a network that runs in discrete time steps (one for each input event), we can translate a pattern that is input sequentially into a pattern of activation across a layer of what we will call "sequence" units, each representing a particular number of time steps before the present.

Binding

Given a static representation of the input sequence, what is required to learn the patterns embodied in the training sequences in Marcus et al.'s experiment? When a listener is presented with a short sequence of syllables, as in the experiment, one of the aspects of the sequence that is salient is the pattern of similarity or differences between the syllables. A device which learns the "rules" implicit in the patterns must be capable of recording this pattern. Thus the task in the experiment involves grouping by similarity, treating two of the syllable words in each three-word sentence as *the same* and the other as *different*. A neural network that simply activated syllable units in sequence, translating this into a static sequence representation, would have no way of performing this grouping.

This grouping of units is one version of a general challenge for connectionist networks known as the **binding problem**. This is the problem of grouping together features during processing and of keeping this information in short-term memory in a form that is accessible and unambiguous. Normally the problem is thought of as one of *segmentation*, binding together features in the input mainly on the basis of proximity in space or time. For vision, segmentation involves dividing a scene into its constituent objects. In the simplest case, each object is represented by a single value on each of a set of sensory/perceptual dimensions (color, texture, etc.), and it coheres in space. For speech, segmentation involves dividing an input stream into its constituent phones, syllables, words, or phrases. Again in the simplest cases, each object can be characterized by a value on a set of dimensions (for syllables, initial consonant, nucleus vowel, etc.), and it coheres in time. It is this sort of grouping that the infants in Saffran et al.'s experiment seem to be doing.

But features in a visual scene or auditory stream may also be grouped on the basis of similarity along some salient dimension. In a scene consisting of an array of blue and red squares, a viewer may treat all of the red squares as one group and all of the blue squares as another, whether or not the squares of the same color appear next to each other and are similar along other dimensions, such as texture. In an auditory stream consisting of sequences of syllables, a listener may group the syllables that resemble each other, whether or not they appear together. It is this kind of grouping that is going on in Marcus et al.'s experiment.

Both of these kinds of grouping or binding involve "belonging together," and we believe that they rely on the same fundamental mechanisms. While most connectionist networks have no way of dealing with the binding problem, some recent models solve the problem through the use of some form of synchronization or alignment (Hummel & Biederman, 1992; Shastri & Ajjanagadde, 1993; Sporns, Gally, Reeke, & Edelman, 1989). Units in the network are outfitted with a dimension of variability in addition to activation, and coincidence along this dimension represents "same object." We will refer to this as the "binding dimension." For Marcus et al.'s task, those sequence units representing positions filled by words treated as the same would be in phase with one another on the binding dimension. It is the sequence units' runtime behavior on the binding dimension that distinguishes the different grammatical patterns from one another.

Relations

But by itself, this other dimension does not suffice. In order to represent the patterns of similarity and difference across sequences of syllables (what Marcus et al. (1999) and Pinker (1999) call "rules"), we need to represent the *relations* between the syllables. In a network with a binding dimension, the relation of sameness can be represented with an excitatory connection, but the relation of difference requires a special mechanism. One way to achieve this is through the addition of explicit **micro-relation units** (**MRUs**) in the network (Gasser & Colunga, 1998). Each MRU represents an association between two features or groups of features which are treated as belonging to different objects or sets of objects. Each MRU has two **micro-roles**, each with its own value along the binding dimension, and when it is highly activated, these values match the values of the feature units that the unit is associated with. To represent a pattern containing three "slots," such as the grammatical patterns in Marcus et al.'s experiment, the network could associate two binary MRUs with one another, mapping the micro-roles in the proper way. Thus for the ABA pattern (*ga ti ga*, *li na li*, *na gu na*, etc.), the network could encode the rule in the form of the connected pair of MRUs shown in Figure 1. While the sort of relational knowledge encoded by MRUs is rudimentary at best, it seems to be all that is needed for the behavior of the infants in Marcus et al.'s experiments.

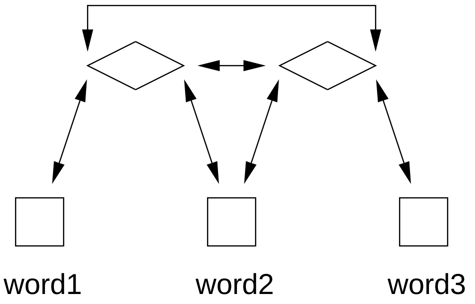

Figure 1: A "rule" (ABA) in the form of a mapping between MRUs, which appear as diamonds. Micro-roles are the ends of the diamonds. The arrows represent positive connections. Each MRU associates one micro-role with each word position. The first MRU represents AB*, the second *AB. The connections between the MRUs map the micro-roles onto each other.

In summary, we believe that a neural network that could simulate the performance of the babies in Marcus et al.'s experiment requires both a means of grouping units together (the binding dimension) and a means of explicitly representing relations (micro-relation units). In the following section, we describe a connectionist network that implements these notions.

The Model

Units

Playpen (Colunga & Gasser, 1998; Gasser & Colunga, 1998) is a neural network architecture of the generalized Hopfield type (Hopfield, 1984; Movellan, 1990) which is designed to represent and learn relational knowledge and to deal with simple sequential patterns. To deal with the binding problem, units vary with respect to their **relative phase angle**, a quantity ranging from 0 to 2π. Each **micro-object unit** (**MOU**), representing an object feature, has a relative phase angle, and when a group of MOUs settles to a state in which they are activated and have similar relative phase angles, the network has implicitly assigned the features represented by those units to a single object in the world. Unlike other networks, Playpen is also outfitted with micro-relation units (MRUs), units which represent primitive binary relations between the features represented by the simpler MOUs. Each MRU has a relative phase angle for each of its two micro-roles. Each micro-role is connected to one or more MOUs, and an MRU is activated to the extent that the groups of MOUs which are positively associated with its two micro-roles are activated and out-of-phase with one another.

Connections

The weights connecting units to each other implement both the simple associations common to all neural networks and the relationships necessary to solve the binding problem. In order to deal with patterns in time, however, each connection has a delay associated with it. The network runs in discrete time, with one time step for each input event (syllable in the case of the simulations reported here). Connections with delay 0 respond in the usual fashion. A connection with delay $d > 0$ causes the unit at the receiving end to respond to the activation that the unit at the sending end had d time steps before. As shown by Kleinfeld (1986) and others, a Hopfield network augmented with delay connections can learn to reproduce the sequences that it is trained on.

As in other neural networks, the sign and magnitude of a weight on a connection has an effect on the activation of the receiving unit. Unlike most other neural networks, the sign and magnitude of the weight also have an effect on the relative phase angle of the receiving unit. Alongside its activation function, each unit has a coupling function which defines this effect. All else being equal, the sending unit *attracts* the phase angle of the receiving unit via a positive connection and *repels* the phase angle of the receiving unit via a negative connection. For connections between MOUs and MRUs, the effect is between the MOU and the micro-role of the MRU. MRUs may also be connected with one another. Because there are two ways in which the micro-roles of two MRUs may pair up, there are two possible patterns of connectivity between each pair of associated MRUs. Again a positive connection implements both excitation and attraction and a negative connection implements both inhibition and repulsion.

Learning

Learning in Playpen, as in most other neural networks, is Hebbian. Because a network may have hidden units, however, simple Hebbian learning often does not suffice; instead **contrastive Hebbian learning** (Movellan, 1990) is used. Learning takes place in two phases, a positive phase in which input units are clamped and learning is Hebbian, and a negative phase in which no units are clamped and learning is anti-Hebbian.[1] When the training patterns have been learned, the two changes cancel each other out because the network's behavior in the two phases is identical.

Network for the Simulation

The network we used for simulating Marcus et al.'s experiments is shown in Figure 2A. There is a Words layer of MOUs, with a single unit for each syllable (word); a Sequence layer of MOUs, with a single unit for each position (lag in time steps behind the current word); and a Grammar layer of MRUs, with a single unit associated with each pair of Sequence units. All of these connections are hard-wired, and there are trainable connections joining the MRUs to one another.

A sequence of words is presented to the network by clamping the activation (but not the phase angle) of the words in succession, one network time step for each word. The hard-wired connections within the Words layer cause words within a sentence consisting of the same syllable to take on the same phase angle and words consisting of different syllables to take on different phase angles. These connections implement the similarity relationships between the different syllables.[2] With no delay, there are negative connections between the different syllable units. With delays of one and two time units, there are positive connections from each unit to itself and negative connections between the different units. With the activations of the Word units clamped, these connections can only affect the units' phase angles, and their effect is to cause similar syllables to have the same phase angle, dissimilar syllables to have opposing phase angles.

The hard-wired connections between the Words and Sequence layers cause the temporal word sequence to be transformed to a spatial representation of the sequence, with one position each for the current and the two preceding words. For the purposes of modeling Marcus et al.'s experiment, we implemented the simplest possible variant of the Sequence layer, one in which there is a single unit for each sequential position. Each of these units is connected to all of the Word units, so it is completely insensitive to the actual word content. The three Sequence units differ only in the delays on their input connections. At the end of an input sentence, all three Sequence units should be activated, and because all are connected by positive connections to the Word units, the Word units will pass on their phase angles to the Sequence units. What remains, then, of the activated sequence of words presented at the Word layer is a non-sequential record of the similarities and differences among the three words. For example, at the end of the sequence *li li na* (right after the network has settled in response to the presentation of *na*), the Sequence units representing the previous word and the second to last word (*li li*) should have the same phase angle, and the Sequence unit representing the current word (*na*) should have the opposite phase angle.

In the Grammar layer, each MRU represents a hypothesis about a relation between two of the Words in a sequence. For a given three-word sequence with two different word types (AAB, ABB, or ABA), we expect two of the MRUs to be activated. Each trainable connection between a pair of Grammar

[1] We consider only the unsupervised version of the learning algorithm here.

[2] We believe it is also possible for the network to learn these connections if presented with many sequences in which syllables are often reduplicated.

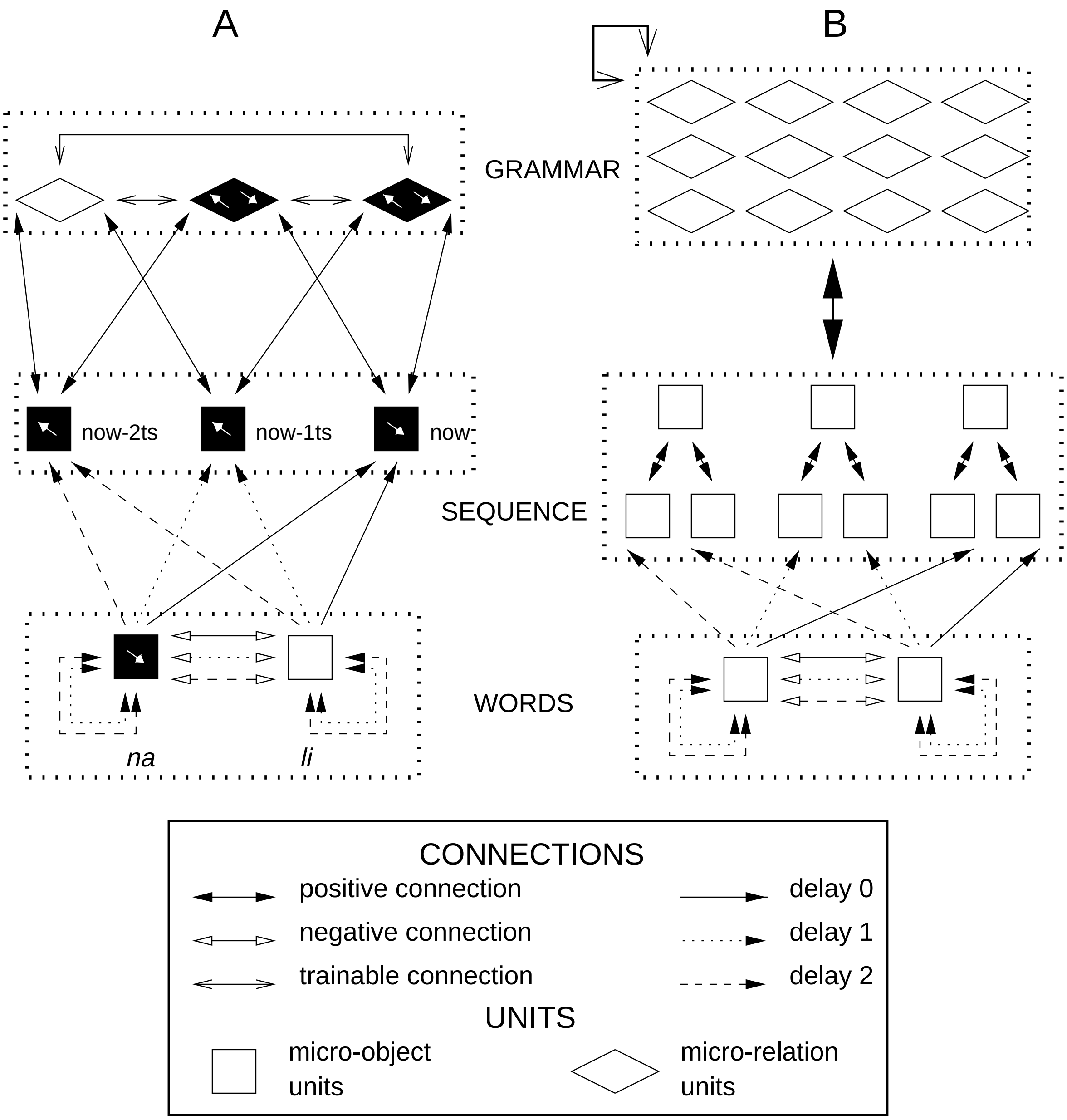

Figure 2: Playpen networks for learning word sequences. (A) Simplified network used in the simulation of Marcus et al.'s experiments. Only two of the four Word units are shown, and the micro-role-to-micro-role mappings are not shown for the MRU connections. The sequence *li li na* has just been presented to the network. MOU and MRU phase angles are represented as arrows. (B) Full-blown version of the network, including word-specific, as well as content-free Sequence units. Not all connections are shown.

MRUs represents a potential correlation between the pairwise word relations. Before training, the Grammar layer has a different attractor state for each of the grammatical patterns. In each of these states, two Grammar units turn on strongly, and the other fails to turn on. Because the hard-wired connections between the Sequence and Grammar layers are relatively weak, however, these attractors are not very reliable, and the Grammar layer may fail to be activated strongly at all or may blend two of the attractors. Following training on a single grammatical pattern, through changes in the weights connecting the Grammar units, we expect the attractor associated with the pattern in the training set to be strengthened and the other two attractors to be weakened. When the network is presented with a sequence of the type it was trained on, it should respond with enhanced activation of the two MRUs that agree with the sequence. On the other hand, when presented with a sequence of either of the other types, it should respond with weakened activation in comparison to its pretraining state. Thus we predict higher overall activation for the Grammar layer for patterns of the type on which the network was trained.

Simulations

We simulated Marcus et al.'s task by training networks on one of the three grammatical patterns: AAB, ABA, or ABB. In each case, the set of training patterns consisted of four different sentences, each formed by randomly combining one-syllable words following the appropriate grammatical pattern. Each network was trained on 50 repetitions of the training set.

The networks were then tested on 12 sentences, four each of the three kinds of grammatical patterns. Each of the test sentences was novel; that is, it was formed by combinations of words that had never been seen before.

Since the Grammar units have learned to be activated by sentences that follow the pattern they were trained on, we expect the activation of the Grammar layer to be higher for sentences that are consistent with this pattern than for sentences that are not. So familiarity with a test pattern is measured in the network as increased activation of the relevant (i.e., trained) units, the Grammar units.

The results from 10 networks trained on each grammatical pattern are shown in Figure 3. The total activation of the Grammar layer was averaged over four trials of each of the test words. The expected interaction between training type and testing type is highly significant ($p < .001$). As shown in Figure 3, the Grammar layer is more activated for novel sentences that follow the grammatical pattern the network was trained on than for novel sentences that follow either of the other two patterns.

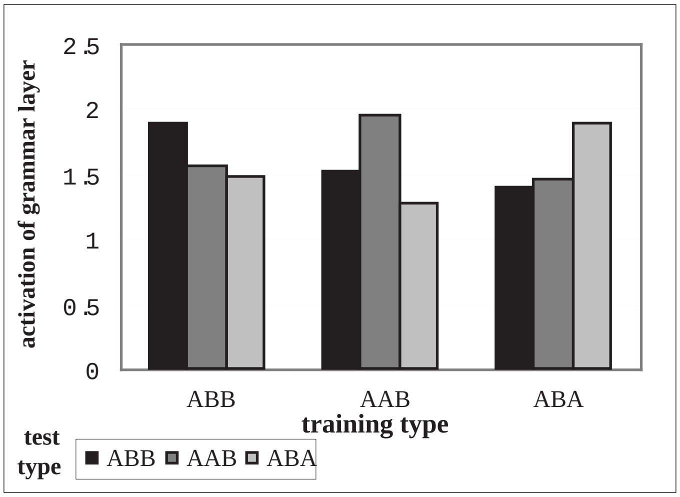

Figure 3: Networks that have been trained on sentences following a certain grammatical pattern respond with more activation to *novel* sentences obeying that same pattern than to novel sentences obeying other patterns.

Conclusions

In this paper we have shown how a neural network with a mechanism for grouping together activated units (relative phase angles) and a mechanism for representing primitive relational knowledge explicitly (micro-relation units) can learn the task of Marcus et al.'s experiments. But as already noted, we were concerned in the simulations only with the simplest sort of network that could model that task. In particular, we included only Sequence units that were completely devoid of content, ones that, through their phase angles, kept a record only of their similarity to the words at other sequence positions. A full-blown network of this type would have Sequence units responding to varying *ranges* of syllables/words, including at the widest end of the spectrum very "abstract" units like those used in the simulation. A simple version of a network of this type is shown in Figure 2B.

We have modeled the task of Saffran et al. using a network similar to that shown in Figure 2B, in which there is a separate sequence unit for each combination of syllable and relative position. This network also has a Word layer consisting of simple units with no phase angles that are fully connected to the sequence units. Trained on sequences of three-syllable words, the Word units come to associate themselves with sequential pairs or triples of particular syllables which recur frequently. Tested on three-syllable sequences later on, the Word units respond more highly to word sequences than to non-word sequences. From the perspective of the network, the Saffran et al. task differs from Marcus et al.'s task mainly in not requiring relative phase angles or MRUs. Because the information that is learned concerns correlations between specific syllables, there is no need for the network to group units through synchronization or to represent explicit relationships between different "objects."[3]

Thus a connectionist network can learn the two tasks, one involving "statistical" knowledge, the other "algebraic rules." While the latter makes use of features of the network that the former does not, the learning algorithm is the same Hebbian algorithm, and the relationships that are learned are implemented through the same sorts of connections. In both cases learning is associationist.

So where does this leave variables? While the Sequence units in the network used to model Marcus et al.'s experiments do function as placeholders in the sense that they stand for relative positions in the input sequence which can be "filled" by different syllables, they are certainly not variables in the usual sense. For one thing, the same three Sequence units are used to represent all three "rules". If the As in AAB are variables, then the Sequence units are not because they can take on phase angles corresponding to AAB, ABB, and ABA. Perhaps the micro-roles of the Grammar MRUs come closer to variables since they force their two phase angles to be distinct. But note that each MRU represents a binary relation which can take part in more than one three-part pattern. It is as if rule fragments are being shared by different rules. Relative phase angles themselves seem to have some of the properties of variables; after all, it is these that are implicitly compared by MRUs. But unlike symbolic variables, phase angles *interact* with one another according to the pattern of connectivity among the units which have the phase angles. And sameness in the network, like similarity in neural networks in general, is graded: two units represent "the same thing" *to the extent* that they are in phase with one another.

But what matters is not whether it is possible to describe

[3] Note that in our simulation of Saffran et al.'s experiment, there is no explicit segmentation of the syllable sequence into separate words. To do this in a Playpen network would require the phase-angle alignment of MOUs associated with each sequence of syllables that corresponds to a word and different phase angles for the MOUs associated with different words. We are currently working on implementing such a network.

the network as having variables. What matters is whether predictions emerge from this model that might differ from those made by a symbolic model. We would make the following predictions:

1. Content should matter, even for "rule"-learning tasks.

   In a full-blown network trained on Marcus et al.'s task (one such as that shown in Figure 2B), Sequence units specific to particular *categories* of syllables would be activated along with the more "abstract" units like those used in the simulation. That is, similarity is still relevant. As a result, we would predict that sentences which overlap in content with those used in training the infants would be treated as more familiar than sentences consisting of completely novel combinations of words. In particular, performance on new instances of the rule should differ from performance on familiar instances of the rule.

2. Difficulty increases with the number of types of syllables in a word.

   Repulsion between three units which are clamped on can lead to three different phase angles, representing three separate "objects," but, depending on the magnitude of the weights connecting the units, there is also an attractor at which there are only two different phase angles. At the same time, MRUs can represent only binary relations, and strong associations between MRUs can only develop for different micro-relations involving the same two objects. Thus Playpen has a strong preference for *two*, and in a four-word version of Marcus et al.'s experiment, we would expect that sequences such as ABCC would be confused with AABB and ABBB. In symbolic models, on the other hand, there is no built-in preference for a particular number of variables.

3. Pairwise relational correlations are always relevant.

   MRUs learn pairwise relational correlations, so the similarity between two "rules" will be determined mainly by the shared pairwise relational correlations rather than any higher-order correlations. In a four-word version of Marcus et al.'s experiment, we would expect ABBC to be treated as more similar to ABBA or ABAC than to ABBB because the first two overlap in 4/5 of the binary relations involving different elements, whereas the third overlaps in only 3/5 of the binary relations.

We have argued that the two kinds of learning exemplified in Saffran et al.'s and the Marcus et al.'s experiments are both associationist. However, a simple neural network with a Hebbian learning rule, even one equipped with delay connections and recurrence to handle time, probably will not suffice to exhibit the variable-like behavior of Marcus et al.'s experiments. We propose that this sort of behavior can be handled by a neural network that is augmented with a mechanism to handle "same thing" and "different thing" and to handle primitive relational knowledge. Thus the Playpen architecture, described in this paper, is a candidate for a general-purpose neural network architecture that can learn both statistical information and rule-like knowledge.

References

Colunga, E. & Gasser, M. (1998). Linguistic relativity and word acquisition: a computational approach. *Annual Conference of the Cognitive Science Society*, *20*, 244–249.

Gasser, M. & Colunga, E. (1998). Where do relations come from?. Tech. rep. 221, Indiana University, Cognitive Science Program, Bloomington, IN.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*, 3088–3092.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proceedings of the National Academy of Science*, *83*, 9469–9473.

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

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 10–17. Morgan Kaufmann, San Mateo, CA.

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by eight-month-old infants. *Science*, *274*, 1926–1928.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations so systematic reasoning: a connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494.

Sporns, O., Gally, J. A., Reeke, G. N., & Edelman, G. M. (1989). Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity. *Proceedings of the National Academy of Sciences*, *86*, 7265–7269.