

Babies, Variables, and Relational Correlations

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 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 can 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 correlational approach to the learning of sequential regularities, and its implementation in a connectionist model, which accommodates both types of learning. We show 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 correlational learning mechanism suffices to perform both of these tasks. Crucially the model makes different predictions than the variable-based account.

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 agreed or disagreed 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, Bandi Rao, and Vishton (1999) presented seven-month-olds with series of three-syllable sequences separated by gaps. Each sequence consisted of two different syllables arranged in a fixed pattern, AAB, ABB, or ABA. For example, in the ABB condition, the presented patterns included sequences such as *le di di* and *ji je je*. Later the infants responded differently to novel sequences of three syllables 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 patterns.

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, not 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 syllable in a sequence 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 connectionist 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 present a model of the learning of regularities in patterns which accommodates both kinds of patterns in terms of **correlations**. We argue that a correlational account, to deal with the tasks in Marcus et al.’s experiment, needs two mechanisms in addition to those usually found in such accounts, neither of which amounts to explicit variables. We show how a connectionist network implementing this theory (the PLAYPEN architecture) can learn aspects of the Saffran et al. task, as well as the Marcus et al. task. What is crucial about this account is not that it handles variable-like behavior within a correlational framework but that it makes predictions that differ from the variable-based account.

Pattern Regularity Learning

Saffran et al.’s and Marcus et al.’s experiments are not directly comparable. In Saffran et al.’s experiments, the boundaries between the patterns must be extracted, while these are provided in Marcus et al.’s task. However, both are learning tasks in which the learner is presented repeatedly with patterns consisting of sequences of syllables and extracts some sort of regularity from the sequences.

We agree with Marcus et al. and Pinker that there are other differences in what is going on in these two tasks, but we believe that both are fundamentally statistical, based on the extraction of **correlations** from input patterns. The main dif-

ference, we argue, lies in what sort of correlations: whether they are content-specific, as in Saffran et al.'s experiments, or relational and based on similarity among the elements within the sequences, as in Marcus et al.'s experiments.

We will consider tasks that are more general than those in the two original sets of infant experiments, what will refer to as **pattern regularity learning**. A learning trial for such a task consists of a pattern (not necessarily auditory) composed of elements arranged in a particular way (either sequentially or spatially), and the regularity consists of tendencies for patterns to resemble each other in particular ways. Resemblances between patterns make reference to the **position** of elements within their patterns, where position may be defined spatially or temporally. Regularity could be concerned only with a single pattern position and not with intra-pattern relationships; for example, *all patterns begin with ba*. But we will only be concerned with regularities that make reference to intra-pattern relationships, as was the case in both sets of infant experiments.

Content-Specific Regularities

In Saffran et al.'s experiments, the resemblances between patterns concern the **specific content** of the patterns. That is, it is particular syllables which are involved in the regularities; certain combinations of syllables tend to recur. The simplest content-specific regularities (other than those that make reference to only a single pattern element) are those involving **pairwise** co-occurrences of specific elements or element features. Examples of such regularities are the following: *ba tends to be followed by gu*; *syllables beginning with b tend to be followed by syllables beginning with g*.

But the regularities in Saffran et al.'s experiments are more complex than these. Rather than simple pairwise regularities, the regularities concern co-occurrences of pairwise co-occurrences. Examples of such **higher-order** regularities are: *when gu is preceded by ba, it tends to be followed by li*; *when a syllable beginning with g is preceded by a syllable beginning with b, it tends to be followed by a syllable beginning with l*.

Not surprisingly, these statistical, content-specific regularities can be handled in a straightforward fashion in connectionist networks. Weights in most connectionist networks represent **correlations** between elements, and the regularities we have been describing are just that. However, correlations between correlations, as in the higher-order regularities, require "handle" units responsible for pairs of particular elements. These handle units can then be joined by connections whose weights encode the higher-order correlations. Figure 1 shows a network of this type. The network is of the attractor (generalized Hopfield) type, and weights are adjusted using the Contrastive Hebbian Learning algorithm (Hopfield, 1984; Movellan, 1990). For simplicity's sake, we assume separate units for the different pattern positions, ignoring the (non-trivial) problem of how element representations are shared across different positions, and we consider only the case of patterns consisting of three elements. Pairwise regularities are represented by strong weights joining pairs of PATTERN units to single CORRELATION units. Higher-order regularities are represented by strong weights on connections joining CORRELATION units. Note that this approach assumes

that higher-order regularities presuppose the pairwise regularities which they are built on. Note also that when there are multiple higher-order regularities, as in Saffran et al.'s experiments, for example, the CORRELATION layer permits these different regularities to be kept separate: one set of units and connections might represent the *ba gu mi* pattern, another the *vi ja lo* pattern.

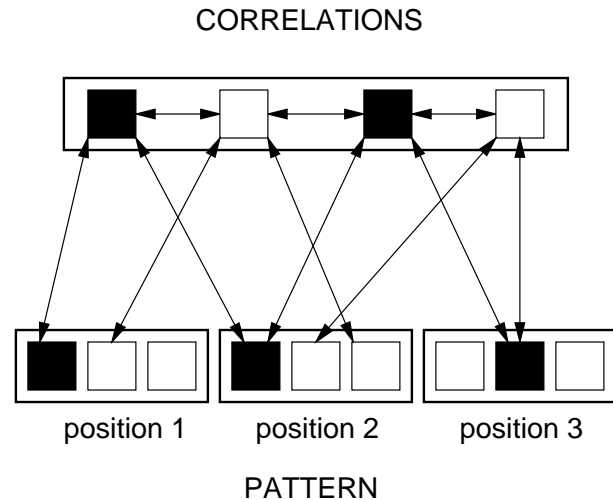


Figure 1: Network for learning content-specific regularities. Only some units and connections are shown.

Just what gets learned by such a network and how it generalizes depend on how the pattern elements are represented. We assume multiple levels of representations differing in coarseness. That is, at the least coarse level, the elements are represented in terms of the largest number of classes; at the most coarse level, they are grouped in terms of a small number of classes. Representations in connectionist networks also differ in the extent to which they are distributed vs. local. Assuming local representations for the sake of simplicity, syllables might be represented at multiple levels of coarseness as shown in Figure 2. Thus the syllable *bis* turns on a unit specific to that syllable, a unit responding to all syllables beginning with *b* and a unit responding to all consonant-vowel-consonant syllables.

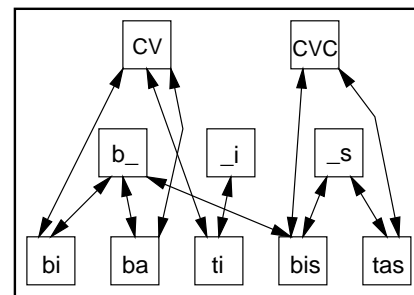


Figure 2: Representation of syllables at multiple levels of coarseness. Only a few units are shown. Arrows represent excitatory connections joining units at different levels of coarseness. Not shown are inhibitory connections forcing winner-take-all at a given level.

Relational Regularities

Alternately, regularity within a set of patterns may be in terms of the similarity of elements within patterns; that is, the regularity may be **relational** rather than content-specific. Again the regularities may be pairwise or higher-order. Examples of pairwise relational regularities are the following: *the first element is the same as the second element; the first element tends to begin with the same consonant as the second element; the first element is different from the second.* Examples of higher-order relational regularities are the following: *the first element tends to be the same as the second element and different from the third; when the first element begins with the same consonant as the second element, the second element has the same vowel as the third.*

In these terms, then, Marcus et al.'s experiments involved both pairwise and higher-order relational regularities, as well as pairwise and higher-order content-specific regularities, though only the relational regularities are reflected in the test items.

In what follows, we discuss how relational regularities, as well as content-specific regularities, are handled within the PLAYPEN architecture.

Accommodating Relational Regularities in a Connectionist Network

Our claim is that relational regularities, like content-specific regularities, are correlations, that is, that they involve statistical patterns of co-occurrence. Further we show how relational correlations can be learned in a connectionist network that differs from more conventional networks in that it has an explicit means of representing and learning about similarity/difference. This requires two augmentations to conventional networks: (1) a second dimension (the "binding" dimension), in addition to activation, along which units vary, and (2) "handle" units which respond to either sameness or difference on the binding dimension.

We view the task presented to the learner in Marcus et al.'s experiments as one of **grouping**, a fundamental aspect of all perceptual processing, both by humans and machines. Presented with a visual or auditory scene, people attempt both to segment it into distinct regions and to group regions together. They segment and group by making use of featural similarity, proximity, and common fate, as well as top-down knowledge of the domain. For segmentation, proximity obviously plays a large role, but for grouping, featural similarity may override proximity. Thus in rhythm perception, where grouping has been studied extensively (Handel, 1989), two elements that are separated by another may be grouped together because of their similarity to each other on some dimension. While segmentation and grouping are in some sense opposing processes, both amount to the **binding** together of regions that would otherwise not be associated with one another.

Thus any cognitive architecture that handles segmentation or grouping must offer a solution to the "binding problem," the problem of how to represent the short-term situation in which distinct cognitive units are treated as part of the "same thing." This problem has been discussed extensively in recent connectionist literature, and a family of related connectionist solutions has been proposed (Shastri & Ajjanagadde, 1993). All of these involve the augmentation of conventional archi-

tectures and algorithms with a further dimension in addition to activation along which processing units can vary. We will refer to this as the "binding dimension." Binding two units then corresponds to coincidence of those two unit's values on the binding dimension. Most often the binding dimension involves the firing of units, and binding itself is synchronization of firing (Hummel & Biederman, 1992; Mozer, Zemel, Behrmann, & Williams, 1992; Shastri & Ajjanagadde, 1993; Sporns, Gally, Reeke, & Edelman, 1989). In PLAYPEN we make use of a simpler approach: alongside its activation, each unit is characterized by an **angle**, ranging from 0 to 2π radians. The particular value taken by a unit's angle is not what is relevant; it is its value relative to that of other units in the network. Units with similar angles are temporarily "bound" together, treated as "the same thing"; units with very different angles (differences close to π radians) are treated as "different things."

To permit the representation and learning of relational correlations, we need one further augmentation. Rather than taking the form of simple connections between units, relational correlations are implemented via "handle" units called **relation units**. These are of two types, **sameness units**, which tend to be activated if their input units are activated and have similar angles, and **difference units**, which tend to be activated if their input units are activated and have different angles. Each of these units represents a pairwise relational correlation of one type or the other, and the connections joining these units represent higher-order relational correlations. Thus the architecture we proposed for learning content-specific correlations (Figure 1) becomes that shown in Figure 3 for relational regularities. Again the network is of the attractor type. We have modified the standard input and activation functions and the Contrastive Hebbian Learning algorithm (Movellan, 1990) to accommodate angles and relation units. For details, see Gasser and Colunga (1998).

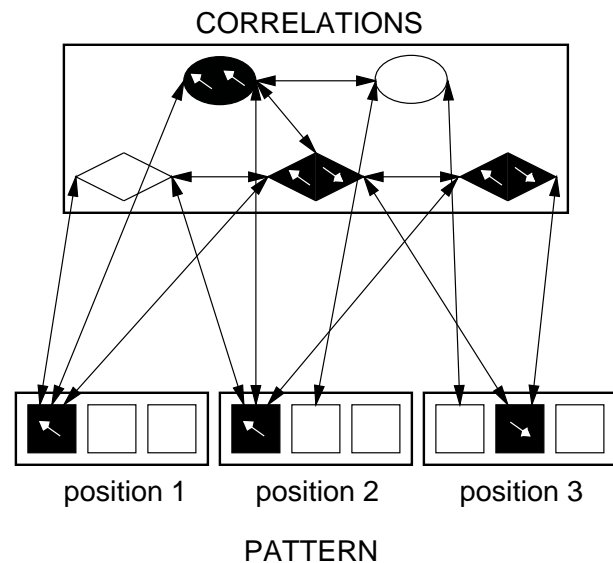


Figure 3: PLAYPEN network for learning relational regularities. Only a few units are shown. Difference relation units appear as diamonds, sameness relation units as ovals. Unit angles are indicated by arrows. A single unit within each pattern position has been activated, leading to the activation of some relation units.

Note that each unit in this network (as in the network in Figure 1) has specific content, but in addition, at any point in time, through its angle, each unit also represents a hypothesis about how the elements in the pattern are to be grouped.

Simulation of Marcus et al.'s Experiment

Now consider again the task of Marcus et al.'s experiment. First, we agree with Seidenberg and Elman (1999) that knowledge about syllable similarity would have been learned prior to the experiment so should already be in place in the architecture. For the PLAYPEN model, this knowledge takes the form of connections (via sameness and difference units) representing the similarity or difference between syllables or syllable features. When the units representing pairs of syllables are clamped in the PATTERN layer, that is, when their activations are fixed at some positive value but their angles are still allowed to vary, these connections cause similar syllables to have the same angle and different syllables to have different angles.

We again assume a range of degrees of coarseness in syllable encodings and, for simplicity, local encodings. The presentation of a pattern, say, *le le di*, takes the form of the clamping of PATTERN units corresponding to these syllables in the relevant sequential positions. Syllable units at greater degrees of coarseness are activated (inhibitory connections between incompatible syllable units prevent all syllable units from being activated as a result of feedback from the coarse units). Further because of the built-in (or previously learned) relational connections implementing similarity, the angles of the syllables take on a pattern representing the *grouping* of the pattern elements: the first two elements make up one group, the third element another. The activated PATTERN units cause particular CORRELATION units to be activated. For example, the difference unit representing *le* in second position and *di* in third position and the difference unit representing some CV syllable in second position and some CV syllable in third position are both activated. Contrastive Hebbian Learning results in the strengthening of connections both into and between the activated CORRELATION units, as well as possibly the weakening of other connections that are not joined by activated units. Figure 4 shows some of the units and connections that are involved.

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

The networks were then tested on 12 sequences, four each of the three kinds of grammatical rules, by clamping the units corresponding to each sequence. Each of the test sequences was novel; that is, it was formed by combinations of syllables that had never been seen before.

Since training the network leads to the strengthening and weakening of connections into and within the CORRELATIONS layer, test patterns should result in more activation on the CORRELATIONS layer if they are consistent with the training set. Thus familiarity with a test pattern was measured in the network as activation of the CORRELATIONS

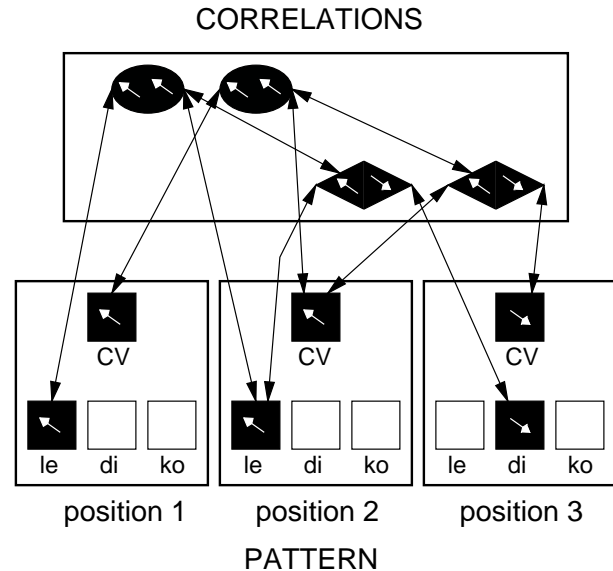


Figure 4: PLAYPEN Network implementing Marcus et al. Only a few units are shown. Connections implementing similarity between PATTERN element units and inhibitory connections between incompatible element units are not shown. The activated (black) units in the PATTERN layer are those that would be active following the presentation of the pattern *le le di*. Four of the relation units that would be activated as a result of this are shown, and ten connections that would be strengthened during the resulting learning. Two of these connections, those joining the units in the CORRELATIONS layer, represent higher-order relational correlations.

units. Because the PATTERN units include very general ones (for example, one that is activated for any CV syllable in second position), the CORRELATIONS layer should be activated relatively highly even by specific syllable sequences it has not been trained on, as long as they are consistent with the training rule.

The average results from 10 networks trained on each grammatical pattern are shown in Figure 5. The total activation of the CORRELATIONS layer was averaged over four trials of each of the test words. The expected interaction between training rule and testing rule is highly significant ($p < .001$). As shown in Figure 5, the CORRELATIONS layer is more activated for novel sequences that follow the grammatical rule the network was trained on than for novel sequences that follow either of the other two rules.

There are several points to note about the way the network learns the tasks.

1. Each unit in the network encodes content information as well as relational information. Thus an activated CORRELATION unit represents at the same time the co-occurrence of particular syllables (or syllable types if it is connected to relatively coarse PATTERN units) and the co-occurrence of syllables bearing a particular similarity relation to one another.
2. Though it cannot perform the segmentation that is a part of Saffran et al.'s task, this network can learn the content-specific correlations in the three-syllable patterns in the task. Since each of the patterns consists of three different syllables, the PATTERN units would take on three differ-

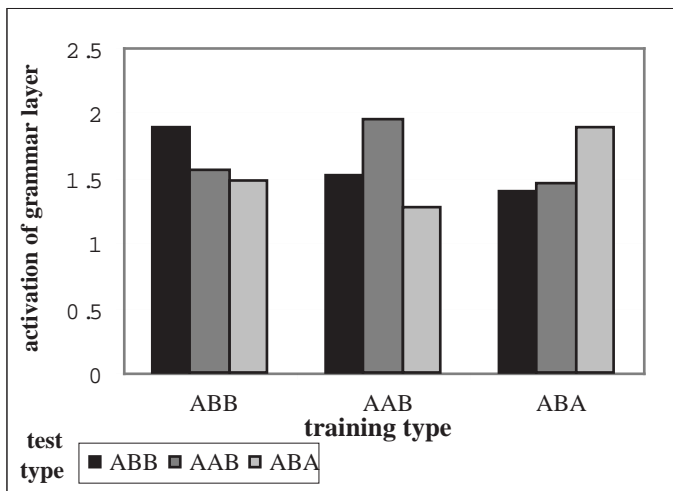


Figure 5: Networks that have been trained on sequences following a certain grammatical pattern respond with more activation to *novel* sequences obeying that same pattern than to novel sequences obeying other patterns.

ent angles for each pattern, activating difference units in the CORRELATIONS layer and resulting in learning on the connections between these units (representing higher-order content-specific regularities).

- While this was not true of Marcus et al.'s task, a set of patterns may embody more than one higher-order relational regularity. For example, in a set of four-element patterns, some patterns might be consistent with the rule AABB and others consistent with the rule ABBA. While we are unaware of experiments testing the ability of subjects to extract such rules, we assume that the ability to learn multiple rules is necessary for language acquisition. A network like that in Figure 4 (but with four positions) could learn both regularities, each as patterns of connections between the six pairwise relational regularities.

Contrasting Two Accounts of Relational Pattern Learning

A Rule-Based Account

A number of models have been proposed to handle the results of Marcus et al.'s experiments (Christiansen & Curtin, 1999; Seidenberg & Elman, 1999). Here we contrast only ours and the rule-based account proposed by Marcus (forthcoming). Marcus argues that tasks such as this one, in fact higher cognition and language generally, rely on the learning and manipulation of **explicit rules** containing **abstract variables**, placeholders that apply to any member of a given class.

Having been trained on a pattern learning task of the type in Marcus et al.'s experiments, the learner extracts an explicit rule of the form AAB, where A and B are now abstract variables in Marcus's sense, and the variables are all associated with some class, say the class of CV syllables (the experiments demonstrate only that infants generalize to other members of this class).

Now consider what patterns will be recognized as familiar after training. Obviously patterns that are identical to those

appearing during training are familiar; if the learner heard the sequence *le le di* during training, that sequence will be recognized later on because it matches the AAB rule. Likewise any pattern consisting of three members of the relevant class for the variables in which the first two elements are identical also matches. So if the relevant class is CV syllables, even if the syllables *ko* and *bi* did not appear during training, the pattern *ko ko bi* will be treated as familiar, apparently just as familiar as *le le di* since all members of the class match the variables equally well. Furthermore, the sequences *le le le* and *ko ko ko* are also familiar since, assuming these variables behave like those in first-order predicate calculus, the rule does not force the third element to be different from the first and second.¹

Now consider what patterns would fail to be treated as familiar. Since identity is all-or-none, patterns in which the first two elements are only similar, such as *le le di* (where ε is the vowel in *bed*) would be treated as unfamiliar. Likewise patterns in which the elements are outside the class over which the variables are defined would not be recognized. Thus, again assuming that CV syllables are the relevant class, *les les dis* would not be seen as familiar.

The Relational Correlation Account

The relational correlation account that we have presented in this paper differs from the rule-based account in that content still matters. This is because, even when what is learned are relational, rather than content-specific, correlations, the correlations apply only to a certain range of elements. The extent of this range depends on the encoding coarseness of the PATTERN units in question, but given a range of degrees of coarseness, we can expect some relatively content-specific relational correlations to be learned, along with some more general relational correlations.

The implication is that the network's response will depend on the degree of similarity between the training and test patterns, as well as on whether the training rule is followed. Patterns that are identical to the training patterns should result in the greatest familiarity. Those that are similar should be treated as less familiar. Those that are quite different, as in Marcus et al.'s experiments, should be still more surprising (though still less so than novel patterns that do not follow the rule).

For the network, the notion of the class over which a variable is defined does not exist. Because CVC syllables share some features with CV syllables, we can expect some generalization to CVC patterns that follow the rule, especially if they share segments with the training syllables.

Further, sameness and difference have equal status in the network, so trained on AAB patterns, the network cannot help but learn that the third element is different from the first and second, as it learns that first and the second are the same. This contrasts with the rule-based approach which requires the learning of an extra predicate to encode the distinctness of the third element.

Finally, difficulty of pattern learning should depend on the number of distinct syllables in the word. When a pattern has

¹Of course, the learner could also extract in addition the explicit constraint that the third element differs from the first and second, but this would seem to be learning "more" than just the rule, so harder or less likely.

three distinct elements, the built-in connections implementing inter-element similarity and difference cause the activated PATTERN units to repel each other's angles, resulting in three different angles. However, depending on the magnitude of the weights connecting the units, there is also an attractor in the network at which there are only two different phase angles. At the same time, relation units can represent only binary relations, and strong associations between relation units can only develop for different relational regularities involving the same two objects (as in Marcus et al.'s experiments). Thus PLAYPEN has a strong preference for *two*, and in a four-syllable 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.

Conclusions and Future Work

In this paper we have shown how a connectionist network with a mechanism for grouping together activated units (angles) and a mechanism for representing primitive relational knowledge explicitly (relation units) can learn the task of Marcus et al.'s experiments. While a PLAYPEN network is perhaps not a conventional neural network, we do not believe it has variables hidden in it. But whether it does or not, the key issue should be whether this model makes different predictions from alternate models, specifically from rule-based models. We have argued in the last section that this is the case. Most of these predictions are testable, and we are currently performing an experiment using visual patterns and adult subjects to test the role of similarity to training patterns in the learning of relational regularities. Preliminary results indicate that subjects are more accurate and faster at judging the familiarity of patterns following the training rule when their content is similar to that in the training patterns, as predicted by our model.

Another potential contribution of our model is the placing of "rule" learning in the context of segmentation and grouping. If we are right, then for auditory patterns such as those in the two sets of infant experiments discussed here, the considerable research on rhythm processing (Handel, 1989) is relevant and should lead to a range of predictions. For example, we might expect the relative timing or loudness of the syllables in patterns to play a role in what is learned.

Relations obviously play a fundamental role in human cognition, and we have argued elsewhere that the relational correlation framework embodied in PLAYPEN accommodates relations without sacrificing the distributed representations and simple Hebbian learning that characterize connectionist networks. Indeed the original motivation for PLAYPEN was the learning of spatial relation terms in language rather than the learning of sequences of syllables. We believe the importance of Marcus et al.'s experiments is not to demonstrate that infants can make use of variables but to show that they are good learners of relational correlations, a capacity that will be crucial as they are exposed to language in all its complexity.

References

Christiansen, M. H. & Curtin, S. L. (1999). The power of statistical learning: no need for algebraic rules. *Proceed-*

ings of the Annual Conference of the Cognitive Science Society, 21, 114–119.

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

Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, Cambridge, MA.

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.

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

Marcus, G. F. (forthcoming). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, Cambridge, MA.

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.

Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation, 4*, 650–665.

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.

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

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to 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.